REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

INTEGRATIVE MACHINE LEARNING APPROACHES FOR ENHANCED CARDIOVASCULAR DISEASE PREDICTION: A COMPARATIVE ANALYSIS OF XGBOOST AND ANFIS ALGORITHMS

Diyar Fadhil Muhyi MUHYI

Master's Thesis

Supervisor

Assoc. Prof. Dr. Oğuz ATA

Istanbul, 2024

INTEGRATIVE MACHINE LEARNING APPROACHES FOR ENHANCED CARDIOVASCULAR DISEASE PREDICTION: A COMPARATIVE ANALYSIS OF XGBOOST AND ANFIS ALGORITHMS

Diyar Fadhil Muhyi MUHYI

Information Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2024

The thesis titled INTEGRATIVE MACHINE LEARNING APPROACHES FOR ENHANCED CARDIOVASCULAR DISEASE PREDICTION: A COMPARATIVE ANALYSIS OF XGBOOST AND ANFIS ALGORITHMS prepared by Diyar Fadhil Muhyi MUHYI and submitted on 13/06/2024 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.

Assoc. Prof. Dr. Oğuz ATA

Supervisor

Thesis Defense Committee Members:

Assoc. Prof. Dr. Oğuz ATA	Department of Software Engineering, Altinbas University	_____
Assoc. Prof. Dr. Sefer KURNAZ	Department of Computer Engineering, Altinbas University	_____
Assoc. Prof. Dr. Aytuğ BOYACI	Department of Computer Engineering, National Defense University	_____

I hereby declare that this thesis meets all format and submission requirements for a master's thesis.

I hereby declare that all information data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Diyar Fadhil Muhyi Muhyi

Signature

DEDICATION

First and foremost, I extend my deepest gratitude to Assoc. Prof. Dr. Oğuz ATA, whose guidance and expertise have been invaluable throughout this research journey. Your dedication and insight have profoundly influenced my scholarly growth and this work.

This thesis is also dedicated to my wonderful husband, whose constant support and encouragement have been my anchor during tough times. Your faith in me has inspired me deeply. To my dear daughter, who has been with me every step of this journey, your laughter and love make every day brighter. I am equally thankful to my parents, whose love and guidance have shaped who I am today. This achievement belongs to all of us. Thank you all for being my strength and support.

ABSTRACT

INTEGRATIVE MACHINE LEARNING APPROACHES FOR ENHANCED CARDIOVASCULAR DISEASE PREDICTION: A COMPARATIVE ANALYSIS OF XGBOOST AND ANFIS ALGORITHMS

MUHYI, Diyar Fadhil Muhyi

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Assoc. Prof. Dr. Oğuz ATA

Date: June / 2024

Pages: 88

Cardiovascular diseases (CVDs) are the leading cause of death globally, underscoring the need for advanced detection and diagnostic methods to enhance patient outcomes. This study investigates the efficacy of two machine learning algorithms, XGBoost and the Adaptive Neuro-Fuzzy Inference System (ANFIS), in predicting heart disease across diverse datasets. Utilizing datasets from the UCI Machine Learning Repository, including Switzerland, Cleveland, Hungarian, Long Beach VA, and Statlog Heart, standard preprocessing techniques such as imputation, standardization, one-hot encoding, and SMOTEENN were applied to ensure consistent modeling conditions. Both models underwent extensive training and optimization. XGBoost excelled, particularly achieving 100% accuracy in the Switzerland and Statlog datasets, while ANFIS demonstrated its strength in modeling complex patterns, notably achieving perfect accuracy in the Cleveland dataset. Performance evaluations using accuracy, precision, recall, F1 score, F2 score, and ROC-AUC score highlighted XGBoost's consistent high precision and recall, vital for reliable CVD diagnosis. In contrast, ANFIS showed potential in clinical settings with its high F2 scores, emphasizing the reduction of false negatives. The study highlights the advantages of using advanced

machine learning models like XGBoost and ANFIS in cardiovascular diagnostics, suggesting further research with larger and more varied datasets to refine these models and advance medical diagnostics using machine learning.

Keywords: Cardiovascular Diseases, Machine Learning, XGBoost, ANFIS, Diagnostic Predictive Modeling.

ÖZET

KARDİYOVASKÜLER HASTALIK TAHMİNİNİN GELİŞTİRİLMESİ İÇİN ENTEGRATİF MAKİNE ÖĞRENMESİ YAKLAŞIMLARI: XGBOOST VE ANFIS ALGORİTMALARININ KARŞILAŞTIRMALI ANALİZİ

MUHYI, Diyar Fadhil Muhyi

Yüksek Lisans, Bilişim Teknolojileri Bilim Dalı, Altınbaş Üniversitesi,

Danışman: Doç. Dr. Oğuz ATA

Tarih: Haziran / 2024

Sayfa: 88

Kardiyovasküler hastalıklar (KVH'ler) dünya genelinde önde gelen ölüm nedenidir ve bu durum, hasta sonuçlarını iyileştirmek için gelişmiş tespit ve tanı yöntemlerine olan ihtiyacı vurgulamaktadır. Bu çalışma, kalp hastalığını tahmin etmede iki makine öğrenmesi algoritmasının, XGBoost ve Adaptif Nöro-Bulanık Çıkarım Sistemi'nin (ANFIS) etkinliğini çeşitli veri setlerinde araştırmaktadır. Çalışmada, UCI Makine Öğrenmesi Veritabanı'ndan alınan İsviçre, Cleveland, Macaristan, Long Beach VA ve Statlog Kalp veri setleri kullanılmış ve modelleme koşullarının tutarlılığını sağlamak amacıyla imputasyon, standardizasyon, one-hot encoding ve SMOTEENN gibi standart ön işleme teknikleri uygulanmıştır. Her iki model de kapsamlı bir şekilde eğitilmiş ve optimize edilmiştir. XGBoost, özellikle İsviçre ve Statlog veri setlerinde %100 doğruluk oranı elde ederek üstün başarı göstermiştir; ANFIS ise karmaşık desenleri modellemedeki gücünü, özellikle Cleveland veri setinde mükemmel doğruluk oranı elde ederek kanıtlamıştır. Doğruluk, kesinlik, geri çağırma, F1 skoru, F2 skoru ve ROC-AUC skoru gibi performans değerlendirmeleri, güvenilir KVH tanısı için XGBoost'un tutarlı yüksek kesinlik ve geri çağırma oranlarını vurgulamıştır. Buna karşın, ANFIS'in yüksek F2 skorları ile klinik

ortamlarda, yanlış negatiflerin azaltılmasına vurgu yaparak, potansiyelini göstermiştir. Çalışma, XGBoost ve ANFIS gibi gelişmiş makine öğrenmesi modellerinin kardiyovasküler tanılarda kullanımının avantajlarını vurgulamakta ve bu modelleri iyileştirmek ve makine öğrenmesini kullanarak tıbbi tanılarda ilerleme kaydetmek amacıyla daha büyük ve daha çeşitli veri setleriyle daha fazla araştırma yapılmasını önermektedir.

Anahtar Kelimeler: Kardiyovasküler Hastalıklar, Makine Öğrenmesi, XGBoost, ANFIS, Tanısal Tahmin Modelleme.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

ANFIS	:	Adaptive Neuro-Fuzzy Inference System
CVDs	:	Cardiovascular Diseases
KNN	:	K-Nearest Neighbors
ML	:	Machine Learning
PCA	:	Principal Component Analysis
RFE	:	Recursive Feature Elimination
ROC Curve	:	Receiver Operating Characteristic Curve
ROC-AUC	:	Receiver Operating Characteristic (ROC) Curve and the Area Under the Curve (AUC)
SMOTEENN	:	SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors)
XGBoost	:	eXtreme Gradient Boosting

1. INTRODUCTION

1.1 BACKGROUND

Cardiovascular diseases (CVDs) continue to be a major global health issue, making a considerable impact on illness and death rates. CVDs are identified by the World Health Organization (WHO) as the primary cause of worldwide deaths, accounting for around 17.9 million fatalities per year. This accounts for 31% of all global deaths (World Health Organization, 2021). The variety of cardiovascular illnesses encompasses a wide range of diseases, such as coronary artery disease, heart failure, arrhythmias, and valvular heart abnormalities. These conditions have a significant impact on healthcare systems and the overall health of society [1].

The manifestation of heart illness is diverse and contingent upon the particular cardiovascular ailment. Typical symptoms consist of angina, which is chest pain or discomfort caused by insufficient oxygen supply to the heart muscle; dyspnea, which is difficulty breathing that is noticeable during physical exertion or even at rest in severe cases; palpitations, which are irregular heartbeats; asthenia or dizziness, especially during physical activity; fatigue, an unusual feeling of tiredness that is not relieved by rest; and edema, which indicates potential heart failure by the accumulation of fluid in the lower limbs [2] – [4].

Heart disease is caused by a combination of hereditary, environmental, and lifestyle factors. Significant risk factors include hypertension, which raises the workload on the heart; hyperlipidemia, which causes changes in the arteries leading to atherosclerosis; tobacco use, which speeds up the development of plaque in the arteries; diabetes mellitus, which is associated with faster progression of atherosclerosis; obesity, which is connected to increased workload on the heart and hypertension; and sedentarism, which is linked to a higher risk of cardiovascular problems [2] – [4].

To achieve an accurate diagnosis of heart failure, it is crucial to use established definitions and gather relevant data, such as the number of hospital admissions linked to heart failure. This highlights the need to detect the disease early on [5]. A range of diagnostic techniques, including electrocardiograms, echocardiography, stress testing, coronary angiography, and blood tests, are crucial for promptly and thoroughly diagnosing heart disease [6].

Moreover, there is variation in cardiovascular disease symptoms between genders, which increases the complexity of diagnoses. For example, whereas chest pain is frequently reported by men, women may also encounter other symptoms such as nausea and profound weariness [7]. Due to the variety, a wide range of diagnostic techniques and competent clinical judgment is required for an appropriate diagnosis [6]. The effectiveness of diagnosing and treating cardiac disease is significantly reduced in settings that do not have access to advanced medical equipment and knowledge [8]. Inadequate diagnostic resources and limited medical competence might result in erroneous diagnoses, while the exorbitant expenses associated with new diagnostic tools can constrain their accessibility [6]. The combined annual expenses related to cardiovascular disease (CVD) and stroke in the United States were estimated to be approximately $351.2 billion for the period of 2014 to 2015. This includes both direct and indirect costs. Direct costs alone increased from $103.5 billion in 1996 to 1997 to $213.8 billion in the same period. These statistics emphasize the significant economic impact of these health conditions [9].

To tackle these difficulties, it is crucial to develop a sophisticated and precise system that can analyze medical data, detect patterns related to heart disease, and forecast impending heart attacks [5]. This method has the potential to revolutionize the management of cardiac disease by allowing for earlier interventions and saving lives [10]. This emphasizes the crucial combination of advanced data analysis technologies with clinical expertise to enhance the prognosis and treatment of cardiovascular disease.

As previously mentioned, diagnosing heart illness may be both costly and time-consuming. To address this issue, Machine Learning (ML) can be utilized. ML plays a vital role in the detection of heart disease by employing sophisticated algorithms to analyze massive quantities of medical data [11]. ML algorithms have the ability to detect intricate patterns and connections in patient data that may not be immediately obvious to human clinicians [12]. ML models can build predictive models for identifying patients at risk of developing heart disease or experiencing cardiac events by leveraging data such as patient demographics, medical history, symptoms, and diagnostic test results [13]. In addition, decision support systems based on ML can aid healthcare practitioners in generating precise and prompt diagnoses by offering insights and recommendations derived from the study of patient data [14]. Integrating ML into the diagnosis of cardiovascular illness has enormous

potential to enhance the effectiveness and precision of diagnostic procedures, ultimately resulting in improved patient outcomes and decreased healthcare expenses [11].

This research focuses on the crucial issue presented by cardiovascular diseases (CVDs), which are a major cause of death and have a considerable impact on the global economy. This thesis utilizes ML techniques to predict cardiac illness by implementing two advanced algorithms, XGBoost and ANFIS. These algorithms facilitate the examination of intricate medical data, assisting in the prompt identification and precise diagnosis of cardiovascular diseases (CVDs). This work highlights the potential of machine learning to revolutionize the healthcare field by combining XGBoost and ANFIS to create advanced diagnostic tools. This not only improves the process of making clinical decisions but also provides a strategic approach to reduce healthcare expenses related to cardiovascular diseases. This research makes a substantial contribution to medical informatics by creating a prediction model that reliably identifies patients who are at risk. This model opens up new possibilities for improving patient outcomes in cardiovascular treatment. In conclusion, this thesis highlights the significance of integrating machine learning with clinical expertise to enhance the diagnosis, treatment, and management of heart disorders. This approach holds the potential to revolutionize healthcare by saving more lives and optimizing the allocation of health system resources.

1.2 STATEMENT OF PROBLEM

The increasing occurrence of cardiovascular diseases (CVDs) on a global scale poses a serious challenge to health systems globally. CVDs are responsible for approximately 17.9 million fatalities each year and have a substantial economic impact. Although medical diagnostics have improved, it is still challenging to detect and diagnose cardiovascular diseases (CVDs) early because these diseases are complicated and have a wide range of symptoms that are impacted by hereditary, environmental, and lifestyle factors. The diversity in symptoms, combined with variations in how symptoms are experienced by different genders, adds complexity to the diagnostic procedure, requiring a range of tests and the expertise of clinical professionals.

In areas where there is limited availability of advanced medical facilities, the difficulty is much more noticeable. Inadequate availability of advanced diagnostic equipment and medical skills can result in incorrect diagnosis, delayed treatment, and higher mortality rates.

Furthermore, the expensive nature of advanced diagnostic technology limits their availability, particularly in areas with little resources, which worsens health inequalities.
This study highlights a significant deficiency in the existing diagnostic approach: the requirement for a novel, economical solution that might improve the precision and promptness of cardiovascular disease detection. Machine learning provides a promising opportunity to fill this gap. By leveraging the capabilities of machine learning to scan large datasets and detect complex patterns, it is possible to create predictive models that can accurately foresee cardiovascular events. Nevertheless, the utilization of machine learning based diagnostic tools for cardiovascular diseases (CVDs) is now at an early stage of development. Therefore, it is crucial to conduct comprehensive research and validation to guarantee their effectiveness and dependability in clinical environments. This study seeks to close this disparity by utilizing the capabilities of XGBoost and ANFIS algorithms, providing a new method to enhance the accuracy and efficiency of diagnosing cardiac problems. This will enable prompt intervention and improve patient outcomes.

1.3 OBJECTIVES AND HYPOTHESIS

1.3.1 The Objective

The main objective of this project is to create and verify a prediction model using machine learning algorithms, specifically XGBoost and ANFIS, to improve the diagnosis and early identification of cardiovascular illnesses (CVDs). The objective of this model is to efficiently evaluate medical data, detecting complex patterns that might forecast the probability of cardiac illness. This, in turn, enables prompt clinical interventions. The research aims to enhance the accuracy of cardiovascular disease (CVD) diagnosis by incorporating sophisticated algorithms. Additionally, it aims to help alleviate the worldwide health and economic burden caused by these diseases. The project seeks to highlight the capacity of machine learning to revolutionize cardiovascular care by providing a scalable and cost-efficient diagnostic tool that can be implemented in various healthcare environments.

1.3.2 The Hypothesis

The hypothesis of this study posits that the incorporation of XGBoost and ANFIS algorithms in the analysis of medical data can enhance the precision and effectiveness of heart disease

prediction, surpassing conventional diagnostic techniques. Our hypothesis posits that the machine learning model would reveal intricate patterns and connections in the data that may not be immediately evident to human clinicians. This, in turn, will enable the model to detect probable heart abnormalities at an earlier stage. The study predicts that the use of a powerful predictive tool will improve clinical decision-making processes. This will lead to better patient outcomes, more efficient allocation of healthcare resources, and a decrease in the cost burden caused by cardiovascular illnesses.

2. LITERATURE REVIEW

2.1 OVERVIEW OF CARDIOVASCULAR DISEASES

Cardiovascular diseases (CVDs) are a collection of conditions that affect the heart and blood arteries, presenting major health challenges worldwide [15]. These diseases are a major concern in the medical profession since they significantly contribute to global rates of illness and death [16]. The range of cardiovascular diseases (CVDs) include several disorders such as coronary artery disease, heart failure, arrhythmias, and valvular heart problems. Each of these conditions has unique pathophysiological characteristics and clinical symptoms [2]. The heart, a vital organ, coordinates the circulation of blood throughout the body, supplying oxygen and nourishment to different tissues and organs. When the health of the cardiovascular system is damaged, it can have far-reaching effects beyond just the heart [17]. It can affect the entire vascular system and, as a result, impact various aspects of human health. The development of cardiovascular diseases (CVDs) is caused by multiple variables, which frequently include an intricate interaction between genetic predispositions, environmental factors, and lifestyle choices [18]. Notable risk factors encompass hypertension, increased cholesterol levels, smoking, diabetes, obesity, and physical inactivity, all of which contribute to the decline of cardiovascular health [9]. On a global scale, the incidence of cardiovascular diseases (CVDs) is very high, affecting millions of individuals who experience the consequences of these illnesses. These problems not only result in a lower quality of life but also place substantial economic burdens on society, including healthcare expenses and reduced productivity [6]. The statistical data demonstrating the influence of cardiovascular diseases (CVDs) highlights the need for improved prevention measures, precise diagnosis techniques, and efficient treatment strategies [18].

Gaining knowledge on the epidemiology and pathophysiology of cardiovascular diseases (CVDs) is essential for creating specific interventions that attempt to decrease the occurrence and seriousness of these illnesses [6]. Healthcare practitioners and researchers can reduce the global burden of cardiovascular illnesses by understanding the underlying mechanisms and risk factors linked with CVDs and developing novel remedies [15].

2.2 DIAGNOSTIC TECHNIQUES FOR CARDIOVASCULAR DISEASE

The diagnostic methods used for cardiovascular disorders are varied and sophisticated, reflecting the nuanced nature of the conditions they are designed to identify and describe [19]. The process of establishing a conclusive diagnosis for cardiovascular disease (CVD) usually begins with a thorough clinical evaluation. During this assessment, healthcare professionals examine the patient's medical history, symptoms, and risk factors. The initial evaluation is of utmost importance since it directs the choice of future diagnostic testing [20].

Physical examinations are essential in healthcare, as practitioners evaluate vital signs, heart sounds, and peripheral circulation to gain information on cardiovascular function [21]. Nevertheless, the intricate characteristics of cardiovascular diseases typically require additional examination using sophisticated diagnostic methods [6].

Blood tests are essential diagnostic techniques that provide biochemical snapshots, enabling the detection of cardiac muscle strain or failure [22]. Biomarkers such as troponins, natriuretic peptides, and lipid profiles offer significant insights into cardiac health and the risk of developing diseases [23].

Imaging tools provide a visual comprehension of the anatomy and physiology of the heart. Echocardiography, employing ultrasound waves, allows for immediate viewing of heart valves, chambers, and contraction patterns, offering a non-intrusive insight into cardiac mechanics [24]. Coronary angiography, which is commonly done after non-invasive testing that indicates a potential problem, offers a comprehensive examination of the status of the coronary arteries. It helps identify any blockages that could lead to heart-related issues [24].

Electrocardiography (ECG) is a fundamental tool in diagnosing heart conditions. It records the electrical activity of the heart to detect irregular heart rhythms, indicators of reduced blood flow to the heart, or evidence of past damage to the heart muscle[25]. Stress testing, in certain instances, can reveal underlying issues that are not apparent while the body is at rest. This type of testing assesses cardiovascular performance when the body is subjected to heightened demands [26].

Although current diagnostic procedures are sophisticated, there are still problems that remain, such as limited accessibility, high costs, and the requirement for expert interpretation of results [27]. Furthermore, the intrusiveness of specific procedures and the possibility of

incorrect results emphasize the necessity for ongoing improvements in diagnostic approaches [28].

To summarize, although the existing diagnostic methods for cardiovascular diseases (CVDs) are strong and varied, continuous innovation and improvement are necessary to increase diagnostic precision, decrease invasiveness, and optimize patient outcomes in the field of cardiovascular care.

2.3 MACHINE LEARNING IN HEALTHCARE

Machine learning, a branch of artificial intelligence, has become a powerful force in healthcare, providing new ways to improve patient care, increase diagnostic precision, and boost treatment effectiveness [29], [30]. Machine learning is centered around creating algorithms that can analyze data, learn from it, and use that knowledge to create predictions or judgments [31]. This allows the algorithms to find patterns and gain insights that may not be easily observable by humans [31].

ML has a wide range of applications in healthcare, including predictive analytics, disease detection [32], and operational efficiencies [29]. These algorithms have the ability to process large amounts of data, such as patient records, imaging, genetic profiles, and clinical notes. They then convert this data into useful information that can be acted upon [33]. An important benefit of machine learning in healthcare is its capacity to process intricate, multidimensional datasets, providing a nuanced comprehension of patient well-being and the advancement of diseases [34]. For example, predictive models have the ability to anticipate the risks faced by individual patients, such as the probability of developing a particular ailment. This allows for proactive interventions and customized care methods [35]. Moreover, machine learning algorithms have the ability to improve diagnostic procedures, as evidenced in fields such as medical imaging [36], where they aid in the identification and categorization of anomalies with more accuracy [30]. These features not only improve the accuracy of diagnosis, but also greatly speed up the process of interpreting results, making it easier to manage patients in a timely manner [29].

Although machine learning has the potential to be integrated into healthcare, there are problems that need to be addressed [31]. Factors like as the protection of data privacy, the transparency of algorithms, and the requirement for thorough validation to guarantee accuracy and dependability are crucial matters to consider [34], [37]. Furthermore, the

achievement of machine learning in healthcare relies on interdisciplinary collaboration, which involves the integration of knowledge and skills from healthcare practitioners, data scientists, and ethicists to effectively negotiate the intricacies of clinical implementation [37], [38].

Machine learning is a very promising field in healthcare that has the potential to reinvent various aspects of patient care, diagnosis, and management. As these technologies advance, they hold the potential to reveal fresh understandings about disease, signaling the arrival of a new age in intelligent healthcare provision.

2.4 MACHINE LEARNING APPLICATIONS FOR CARDIOVASCULAR DISEASE PREDICTION

The application of machine learning in predicting cardiovascular disease is a burgeoning field, evidenced by a multitude of studies aiming to enhance diagnostic accuracy through innovative algorithms. A study [39] introduced a modified differential evolution algorithm, DE/rand/2-wt/exp, with the aim of enhancing feature selection for predicting cardiovascular disease. The improved technique was employed in a model that integrated the Fuzzy Analytic Hierarchy Process with Artificial Neural Networks, resulting in an accuracy rate of 83% in predicting heart disease. The study emphasizes the promise of this integrated method in enhancing predictive analytics in healthcare, however, it is advisable to further validate it with larger datasets.

In a comprehensive analysis [40], the efficacy of six different machine learning algorithms was examined in the context of cardiac disease prediction. Evaluation methods include logistic regression, SVM, ANN, KNN, decision trees, and Naïve Bayes. Logistic regression was the most accurate, with 85% accuracy and high sensitivity and specificity. The study highlights logistic regression and ANN's ability to predict cardiac disease. SVMs excel in finding positive cases. The findings suggest that machine learning could considerably enhance cardiac illness diagnosis. To corroborate these conclusions, larger datasets need be studied.

Another innovative study [41] enhanced heart disease diagnosis using SVM classifiers, employing Fisher scores and Matthews correlation coefficients are used to select the best characteristics. The method was used on Cleveland, Hungarian, Switzerland, and SPECTF

UCI datasets. For each dataset, accuracy increased by 81.19%, 84.52%, 92.68%, and 82.7%. This shows how well the method improves heart disease prediction.

Research [42] explored methods to refine heart disease predictions by identifying key variables and employing various data mining approaches. The Cleveland dataset was used to evaluate seven classification methods and uncover nine predictive characteristics, including gender and chest pain kind. Vote, a hybrid data mining algorithm combining Naïve Bayes and Logistic Regression, accurately predicts heart disease with 87.4% accuracy. This emphasizes the need of choosing characteristics and methodologies carefully when constructing prediction models.

A study [43] aimed to elevate cardiovascular illness prediction accuracy by integrating machine learning techniques, Relief and LASSO feature selection combined with RFBM hybrid classifiers. A dataset from five sources is analyzed and evaluated using many criteria. It concludes that RFBM with Relief feature selection has 99.05% accuracy, promising early disease detection and healthcare advancements.

The HRFLM method [44], employing a diverse array of machine learning techniques on the Cleveland UCI dataset, fully leveraged available features without restrictions, attaining an impressive 88.7% accuracy in heart disease prediction. This highlights HRFLM's capability as an effective diagnostic tool.

An advanced methodology [45] utilizing an optimized XGBoost classifier, enhanced through Bayesian optimization and One-Hot encoding, outperformed standard classifiers with a 91.8% accuracy on the Cleveland dataset. This underscores the benefit of meticulous hyper-parameter tuning in enhancing model efficacy.

The MIFH framework [46] utilizes Factor Analysis of Mixed Data (FAMD) alongside various machine learning techniques, significantly improving heart disease prediction accuracy through the UCI Cleveland dataset, particularly when implementing the Random Forest model, which achieved a 93.44% accuracy rate. This emphasizes the advantages of combining advanced feature extraction techniques with machine learning to boost detection accuracy.

The integration of IoT with deep learning [14] through a system utilizing a Deep Learning Modified Neural Network (DLMNN) for patient monitoring and heart disease prediction showed a remarkable 96.8% accuracy, surpassing traditional algorithms and illustrating the potential synergy between IoT and deep learning in medical diagnostics.

A novel framework [47] combining Support Vector Machine (SVM) with fuzzy logic for decision-level fusion demonstrated a 96.23% accuracy in cardiac illness prediction, illustrating a significant enhancement over existing methods. This two-step process, beginning with supervised machine learning and followed by fuzzy logic-based decision fusion, offers a comprehensive approach to predicting heart disease.

The High-Dimensional Partitioning Method (HDPM) [48], integrating DBSCAN for outlier detection, SMOTE-ENN for data balancing, and XGBoost for prediction, showcased its effectiveness on the Statlog and Cleveland datasets with accuracies of 95.90% and 98.40%, respectively. This indicates its superiority in diagnosing heart disease compared to earlier models.

A health monitoring system [49] combining IoT with the Random Forest algorithm for disease prediction underscored the algorithm's efficacy, achieving up to 97.26% accuracy in disease diagnostics. This showcases the transformative potential of leveraging IoT and machine learning in healthcare.

An approach [50] combining ensemble deep learning with feature fusion for cardiac illness prediction analyzed data from sensors and EMRs, achieving a 98.5% accuracy rate, highlighting its potential in improving healthcare outcomes through advanced data analysis and deep learning.

The Enhanced Deep Learning Assisted Convolutional Neural Network (EDCNN) [51], applied on the IoMT platform, leverages deep learning and advanced mathematical modelling to analyze patient data, achieving a diagnostic accuracy rate of up to 99.1% in heart conditions. This indicates a promising future for cloud-based medical diagnostics.

In another innovative study [52], a diagnosis system employing Modified Salp Swarm Optimization (MSSO) and Adaptive Neuro-Fuzzy Inference System (ANFIS) within the IoMT framework showed substantial effectiveness in heart disease diagnosis, achieving an accuracy of 99.45% and precision of 96.54%. This underscores the potential of integrating advanced optimization and fuzzy logic techniques in enhancing medical diagnostics in the digital age.

These studies collectively highlight the dynamic and potent role of ML in cardiovascular disease prediction, offering insights into its application for more accurate, efficient, and timely diagnostics, which could revolutionize patient care and healthcare systems worldwide.

2.5 THEORETICAL FRAMEWORK OF XGBOOST AND ANFIS ALGORITHMS

The growth of machine learning has brought about intricate algorithms that can offer a substantial understanding of intricate datasets. Notably, the Extreme Gradient Boosting (XGBoost) and Adaptive Neuro-Fuzzy Inference System (ANFIS) are particularly remarkable for their distinct capacities, particularly in the realm of cardiovascular disease prediction.

2.5.1 XGBoost Algorithm

XGBoost is a sophisticated implementation of gradient boosting machines, created to be extremely efficient, adaptable, and transferable [53]. The system functions based on the idea of gradient boosting, which entails constructing a model in the shape of a collection of weak prediction models, commonly referred to as decision trees [54]. The fundamental concept is to progressively incorporate predictors into an ensemble, with each one rectifying the errors of its predecessor, thus enhancing the accuracy of the model [53]. XGBoost stands out by including many optimization strategies that improve performance and speed. These include a novel tree-learning algorithm that effectively handles sparse data and a scalable, distributed computing framework that speeds up computations [53].

The algorithm's efficacy in utilizing system resources and its capacity for parallel computing make it a solid tool for addressing extensive and intricate datasets [53]. In addition, XGBoost incorporates regularization settings to mitigate overfitting, a prevalent issue in machine learning models. This ensures that the model maintains its ability to generalize to new, unknown data [45].

2.5.2 ANFIS Algorithm

The Adaptive Neuro-Fuzzy Inference System (ANFIS) is a hybrid intelligent system that combines the learning capabilities of neural networks with the knowledge representation of fuzzy logic in order to accurately model complex nonlinear functions [55]. ANFIS combines the principles of neural networks with fuzzy inference systems by utilizing a neural network structure to execute a fuzzy inference system, thereby taking use of the advantages offered by both methodologies [56].

ANFIS consists of five layers: fuzzification layer, rules layer, normalization layer, defuzzification layer, and output layer. The fuzzification layer converts input values into fuzzy membership values. The rules layer applies fuzzy logic operations. The normalization layer calculates the ratio of each rule's firing strength to the sum of all rules' firing strengths. The defuzzification layer generates a crisp output for each rule. The summation layer computes the overall output as the weighted average of all rule outputs [57].

ANFIS is capable of accurately approximating nonlinear functions, making it well-suited for jobs involving complex and poorly understood relationships between input and output variables [55]. The versatility and learning capabilities of ANFIS make it a highly useful tool for predictive modelling in diverse sectors, such as healthcare.

2.5.3 Integration in Cardiovascular Disease Prediction

The theoretical foundations of XGBoost and ANFIS offer a strong platform for tackling the difficulties associated with predicting cardiovascular illness. The combination of XGBoost's efficacy in managing extensive datasets and its robust classification capabilities, along with ANFIS's capacity to represent non-linear connections through adaptive learning, provides a comprehensive methodology for comprehending and forecasting the risk of heart disease. By utilizing these algorithms, researchers may create prediction models that are both precise and capable of revealing complex patterns and connections within medical data. This eventually enhances diagnostic procedures and patient results in cardiovascular care.

3. METHODOLOGY

3.1 RESEARCH DESIGN

This study adopts quantitative research methodology, deploying advanced machine learning algorithms to enhance predictions concerning cardiovascular diseases (CVDs). Central to this investigation is the exploration of the predictive power harnessed by machine learning models, specifically XGBoost and ANFIS, to refine the accuracy of CVD diagnosis. The study methodically structures data pattern analysis and model performance evaluation through quantifiable metrics, embracing a data-driven approach to medical diagnostics.

Commencing with Figure 3.1, the research delineates the XGBoost model's operational pathway. This begins with an essential phase of data preprocessing, which includes missing value imputation using KNN, data normalization, and categorical variable transformation via One-Hot Encoding. To address the challenge of class imbalances prevalent in medical datasets, SMOTEENN is employed to ensure a balanced representation of classes. In the feature engineering step, PCA is applied to reduce dimensionality and highlight significant predictive features, followed by a selection process to identify the most informative predictors. The model is then fine-tuned through hyperparameter optimization, culminating in a robust model training and evaluation stage.

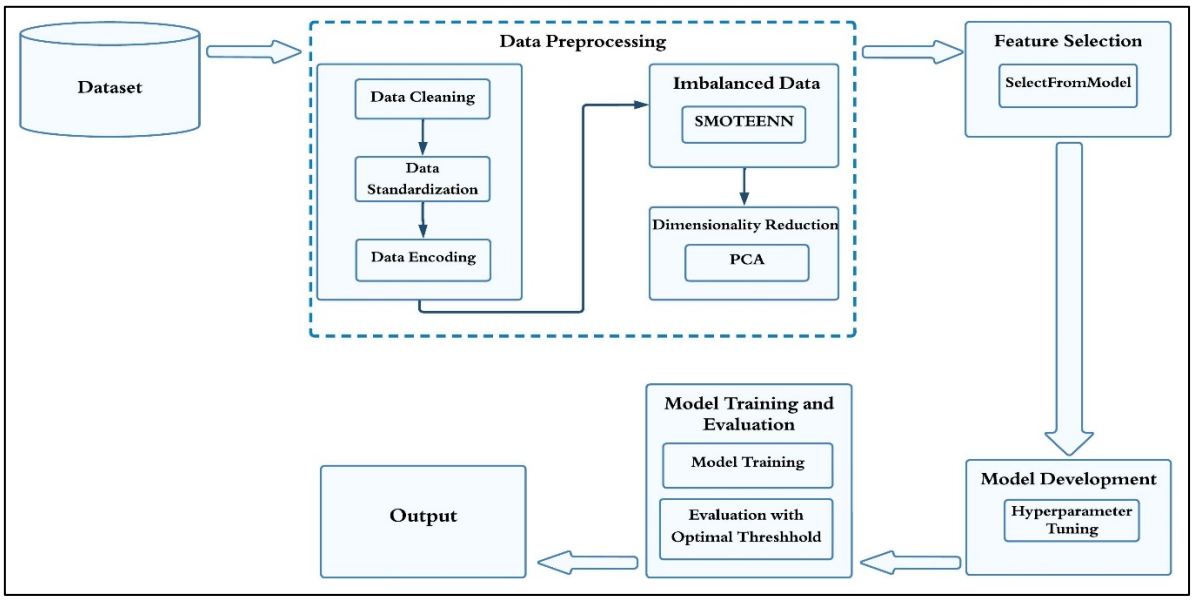

Figure 3.1: XGBoost Model Research Framework.

Post XGBoost analysis, the study mirrors these initial steps in the ANFIS model's framework, outlined in Figure 3.2. The data undergoes similar preprocessing and balancing techniques, ensuring a consistent and equitable dataset for subsequent modeling. The ANFIS model leverages a distinct set of hyperparameters, reflecting its unique computational architecture that combines fuzzy logic with neural network adaptability.

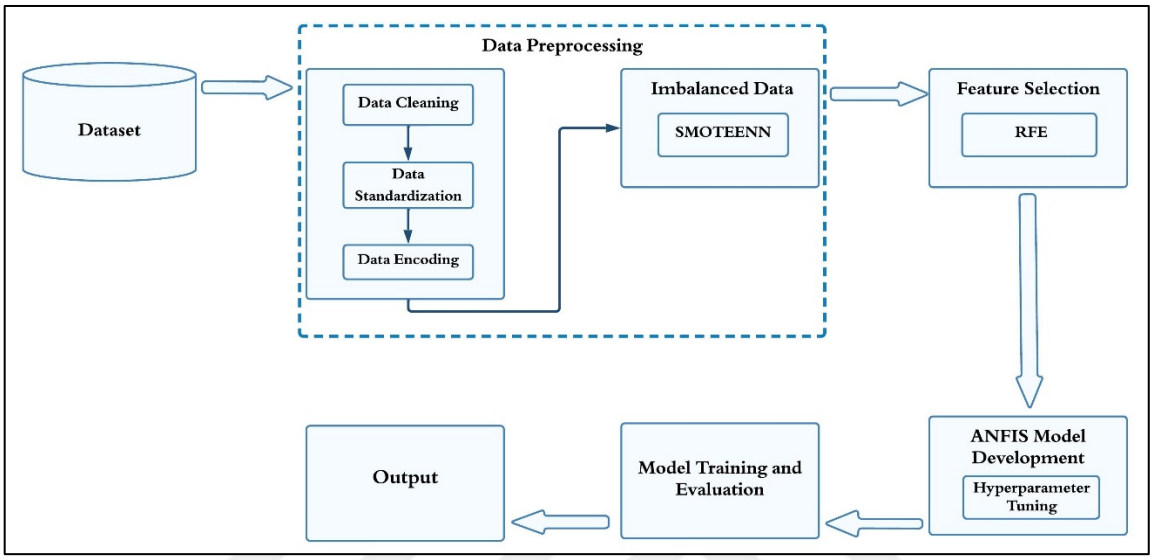

Figure 3.2: ANFIS Model Research Framework.

By embracing five comprehensive datasets—Switzerland, Cleveland, Hungarian, Long Beach VA, and Statlog Heart—the study encompasses a wide spectrum of CVD profiles. This diversity underpins the robustness of the research, ensuring its applicability across varied patient demographics and clinical scenarios.

Sequential phases of the research design are meticulously crafted, underpinning the collection and preprocessing of high-quality, homogenous data, thereby priming it for effective analysis. The ensuing phase implements the XGBoost and ANFIS algorithms with an emphasis on precision-tuning to the task at hand. The performance of these models is scrutinized through established benchmarks that quantify their diagnostic accuracy, sensitivity, and specificity for CVD.

This project is devoted to unearthing the latent capabilities of ML in the diagnostic realm of cardiovascular conditions. Through a structured, systematic, and quantitative framework, this research is poised to unearth valuable insights for the medical field. The expected outcomes are envisaged to influence future clinical procedures and inform policy development in cardiovascular healthcare management.

3.2 DATA COLLECTION

This study utilizes data from five reputable datasets, obtained from the UCI Machine Learning Repository, to create a comprehensive analysis framework for predicting cardiovascular disease. The datasets were carefully selected to include a diverse range of patient demographics, clinical symptoms, and diagnostic outcomes, thereby creating a complete foundation for the development and validation of prediction models.

The selection includes:

Cleveland is well-known for its extensive collection of clinical data, which includes diagnostic test results and signs of heart disease [58].

The Hungarian provides comprehensive data from the Hungarian Institute of Cardiology, located in Budapest, which enriches the diversity of the study[58].

Switzerland is provided by the University Hospital in Zurich, Switzerland, focuses on distinct cardiovascular problems that are distinctive to the region [58].

The Long Beach VA integrates data from veterans to get insights into the cardiovascular health concerns experienced by this group [58].

The Statlog dataset is widely recognized for its role in enabling statistical analyses and machine learning applications in the field of cardiovascular research [59].

The acquisition process from the UCI Machine Learning Repository involved verifying the integrity and relevance of each dataset to confirm its suitability for developing advanced machine-learning models for predicting cardiovascular disease. The following Table 3.1 illustrates the description for each dataset.

Table 3.1: Datasets Description.

Dataset Name	No. Features	No. Sample
Cleveland	14	304
Hungarian	14	295
Switzerland	14	124
Long Beach VA	14	201
Statlog	14	271

3.3 DATA PREPROCESSING

Data preparation is a crucial step in assuring the efficacy of machine learning models, especially when working with intricate datasets for the prediction of cardiovascular disease. This phase encompassed a sequence of methodical techniques aimed at refining the raw data obtained from the Switzerland, Cleveland, Hungarian, Long Beach VA, and Statlog Heart datasets. The objective was to improve the quality and compatibility of the data before utilizing the XGBoost and ANFIS algorithms for analysis.

3.3.1 Data Cleaning

Data cleaning frequently entails dealing with missing data in datasets, which is a crucial stage for ensuring precise data analysis or machine learning [60]. The K-Nearest Neighbors (KNN) imputation approach provides a solution by utilizing the similarity between data points. This method addresses missing values by identifying the 'k' nearest neighbors using a distance measure and calculates the missing value as an aggregation of these neighbors' values [61].

For a dataset X with a missing feature value x_{miss} KNN imputation computes the imputed value, $\hat{x}_{miss}$, as follows by equation (3.1):

$$\hat{x}_{miss} = \frac{1}{k} \sum_{i=1}^{k} w_i \times x_{(i,miss)} \tag{3.1}$$

In this context, $N = \{n_1, n_2, \dots, n_k\}$ represents the set of 'k' nearest neighbors. The value of the missing feature for the i^{th} neighbor is denoted as $x_{i,miss}$, and w_i indicates the weight assigned to the contribution of the i^{th}neighbor. When using uniform weights, each neighbor's contribution is the same ($w_i = 1$ for all i), resulting in the imputed value being the average of the neighbors' values for the missing feature [61].

The selection of 'k', or the number of neighbors, has a substantial impact on the quality of the imputation. Inadequate number of neighbors may fail to collect the essential information, whilst an excessive number could potentially create unwanted noise [61]. The optimal value of 'k' is usually established by doing empirical evaluation that is specific to the dataset.

KNN imputation relies on the assumption that data points that are proximate in feature space exhibit comparable values [61]. This makes it a suitable method for our datasets, where this proximity may reliably anticipate missing values. The effectiveness of the strategy depends

on the careful selection of 'k' and the distance metric, with the goal of effectively utilizing the inherent data structure [61].

3.3.2 Data Standardization

Data standardization is an essential procedure in data preprocessing, particularly for machine learning [62]. The process entails converting the characteristics of a dataset to have an average of zero and a standard deviation of one [62]. This modification guarantees equitable contribution from every feature, enhancing the efficiency and efficacy of numerous algorithms.

The equation (3.2) represents the process of standardizing a feature value $\boldsymbol{x}$ in a dataset.

$$x_{std} = \frac{x - \mu}{\sigma} \tag{3.2}$$

Assume that $\boldsymbol{x}$ represent the original value of a feature, $\boldsymbol{\mu}$ represent the mean of the feature, and $\boldsymbol{\sigma}$ represent the standard deviation of the feature. By applying this method to each feature, the dataset is normalized, which means that the features will have an average($\boldsymbol{\mu}$) of 0 and a standard deviation ($\boldsymbol{\sigma}$) of 1 [63].

Normalization is crucial for algorithms that are affected by the magnitude of data or when features cover a wide range of scales [63]. Data standardization is an essential step in many data analysis and machine learning pipelines. We used data standardization to make sure all the features were dimensionless and scaled equally. This process helps to improve the effectiveness of model training and analysis [46].

3.3.3 Data Encoding

Efficiently handling categorical data is essential in machine learning as the majority of algorithms require numerical input. One-Hot Encoding is a crucial preprocessing technique that converts variables into a binary matrix, making the data suitable for machine learning models [64].

One-Hot Encoding is a technique in machine learning that turns a categorical feature X, which has m distinct categories, into a binary vector of size m. The mathematical representation of this procedure is as follows:

$$x_i = \begin{cases} 1, & if\ X\ category_i \\ 0, & otherwise \end{cases}$$

for $i = 1, 2, \ldots, m$ where x_i signifies the presence (1) or absence (0) of the i^{th}category in an observation [65].

This method ensures appropriate representation of categorical variables without assuming any undesired order, which is essential for accurately and efficiently training and predicting our models [64].

3.3.4 Imbalanced Data

Imbalanced data in machine learning refers to situations where the distribution of classes is uneven, resulting in a bias towards the majority class in the model. SMOTEENN is a technique that combines Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN). It solves the problem by creating artificial examples of the minority class and eliminating incorrectly categorized instances of the majority class [66]. In the SMOTE algorithm, a new synthetic instance $x_{new}$is generated by interpolating between a minority class instance a and its nearest neighbor b as the following equation (3.3).

$$x_{new} = x_a + \lambda \cdot (x_b - x_a) \tag{3.3}$$

where x_a and x_b are the feature vectors of a and b, respectively, and λ is a random number between 0 and 1 [67].

We employed this technique to augment the representation of the minority class, while ENN aids in refining the dataset by removing noisy occurrences, resulting in a more balanced and dependable dataset for training machine learning models [66].

3.3.5 Dimensionality Reduction

For our research, we utilized principal Component Analysis (PCA) to reduce the dimensionality of our datasets [68]. This method is essential for simplifying the data by translating the original variables into principal components [60]. PCA does this by computing the eigenvectors and eigenvalues of the covariance matrix of the data. The eigenvectors determine the new axes, known as principal components, while the eigenvalues quantify the amount of variation explained by each component [68].

Mathematically, the procedure entails obtaining the covariance matrix from the dataset X, followed by extracting its eigenvectors and eigenvalues [68]. The principal components are

subsequently arranged in descending order based on their eigenvalues, which indicate the amount of variation captured by each component [60]. By utilizing the 'mle' method, we can automatically optimize the selection of the most important characteristics of the data [69]. Specifically, we select the top m principal components, where m is determined by this optimization process. This approach ensures that we keep the most relevant elements of the data [60].

The deliberate decrease in dimensionality, as implemented in our study, helps to address the problem of high dimensionality, improve computational performance, and preserve the integrity of the dataset's informational content.

3.3.6 Data Splitting

In machine learning, the process is to split the dataset into separate training and testing sets. This is done in order to train the model using the training set and then evaluate its performance using the testing set [70]. The ‘train_test_split’ function partitions the data into a test set and a training set, with a defined percentage (e.g., 30%) allocated to the test set and the remaining portion assigned to the training set. The separation of data into distinct sets enables impartial evaluation of the model by testing it on unseen data, hence verifying the model's capacity to apply to new situations and avoiding it from overfitting the training data. The study provides a strong groundwork for the application of machine learning algorithms to predict cardiovascular illness, by carefully preprocessing the data. The preparatory steps improved the quality of the input data and refined the settings for training and testing the XGBoost and ANFIS models. This ensured that the analysis was robust, and the results were reliable.

3.4 FEATURE SELECTION TECHNIQUES

3.4.1 Feature Selection with SelectFromModel in XGBoost

Within the framework of XGBoost, feature selection is simplified by utilizing the inherent computation of feature importance during the model training procedure. SelectFromModel is a meta-transformer that selects features based on their determined importance. The threshold for preserving features is set to the median value, meaning that only features with importance over this value will be kept [71].

The significance of a feature in XGBoost is typically measured mathematically using measures like as gain. Gain is the average enhancement in accuracy that a feature brings across all trees. The equation (3.4) shown below represents the method for estimating the significance I_j of a property j :

$$I_j = \frac{\sum_{tree} Gain_j}{\sum_{tree} Splits_j} \tag{3.4}$$

where $Gain_j$ is the gain of feature j summed over all trees, and $Splits_j$ is the number of times feature j is used to split the data [72].

During feature selection, only the features with relevance scores that are equal to, or more than the median are kept [71]. This approach focuses on the most important predictors, which might potentially enhance the performance of the model [73].

3.4.2 Feature Selection with Recursive Feature Elimination in ANFIS

The Recursive Feature Elimination (RFE) plays an important part in improving the Adaptive Neuro-Fuzzy Inference System (ANFIS) for cardiovascular disease prediction. RFE employs a methodical approach to identify the most crucial features, hence improving the predicted accuracy of the model. This approach progressively improves the feature set by selecting only the most important features, resulting in a more efficient model with reduced input dimensions [74].

The (RFE) approach begins with a whole set of features and gradually removes the least significant ones, taking into account their impact on the performance of the model. The iterative procedure continues until a predetermined number of important features are picked, which are considered to be the most advantageous for improving the accuracy of the model [75].

Utilizing (RFE) with a machine learning algorithm such as XGBoost, which is recognized for its accurate feature importance ranking, enables the discovery of these crucial characteristics. The selection process is of utmost importance as it has a direct impact on the efficiency and usefulness of the ANFIS model in predicting heart disease. By prioritizing the most useful characteristics, the model not only becomes more effective in terms of computational resources but also improves its ability to make accurate predictions. This reduces the chances of overfitting and enhances its performance on new and unexplored data [75].

3.5 IMPLEMENTATION OF MACHINE LEARNING ALGORITHMS

3.5.1 XGBoost Algorithm Implementation

XGBoost, also known as Extreme Gradient Boosting, is a prominent machine learning algorithm that is highly regarded for its proficiency in analyzing structured data and its ability to accurately predict outcomes. The system functions by utilizing a boosting process that systematically rectifies errors from previous decision trees [45], which proves to be highly efficient in the intricate domain of cardiovascular disease prediction. The effectiveness of XGBoost can be due to its exceptional performance in classifying tasks, its ability to handle tabular data effectively, and its skill in managing non-linear connections between features. These qualities make it an ideal choice for medical predictive modelling [72], [75].

The algorithm's effectiveness depends on the careful adjustment of hyperparameters such as the number of trees (n_estimators), tree depth (max_depth), learning rate (η), and node split regulation (γ). The parameters are carefully tuned, usually through cross-validation methods like GridSearchCV, to improve the model's capability to identify intricate patterns and achieve accurate predictions [45]

The core of XGBoost's model training revolves around minimizing a loss function [72] as the following equation (3.5).

$$L(\theta) = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_k) \tag{3.5}$$

This approach combines the prediction error with a regularization term in order to reduce overfitting. The equation (3.6) is used to maximize the gain from each decision tree split [72].

$$G = \frac{1}{2}\left[\frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} + \frac{(\sum_{i \in R} g_i)^2}{\sum_{i \in R} h_i + \lambda} - \frac{(\sum_{i \in Node} g_i)^2}{\sum_{i \in Node} h_i + \lambda}\right] - \gamma \tag{3.6}$$

The mathematical methodology employed by XGBoost guarantees the capture of both the fundamental data patterns and the necessary precision for medical diagnostics. Using XGBoost for cardiovascular disease prediction demonstrates its analytical capabilities in addressing the intricacies of medical data and providing a dependable foundation for clinical decision-making [45].

3.5.2 ANFIS Algorithm Implementation

3.5.2.1 ANFIS architecture

The fuzzy logic integration with neural network principles makes up the ANFIS framework, which provides a systematic approach for data analysis and modelling. The architecture and mathematical operations of ANFIS are as follows [57]:

a. Fuzzification Layer

In this layer, numerical inputs are converted into fuzzy values through membership functions which are commonly shaped as sigmoid. The membership value of input is formulated according to the following equation (3.7):

$$\mu A_{ij}\ (x_i) = \frac{1}{1 + exp\left(-\ b_{ij} \cdot (x_i - a_{ij})\right)} \tag{3.7}$$

where a_{ij} and b_{ij} represent the center and width of the sigmoid function, respectively.

b. Rule Application Layer

This layer applies the fuzzy logic rules to the fuzzified inputs and generates preliminary rule outputs. The firing strength w_j of each rule is determined, which is a combination of the membership values of the inputs associated with that rule. Normally, this is calculated using an AND operation described in the following equation (3.8)

$$w_j = \prod_{i=1}^{n} \mu A_{ij}(x_i) \tag{3.8}$$

where n is the number of inputs.

c. Normalization Layer

This layer ensures that the firing strengths of the rules are normalized to total one. Therefore, the firing strength of every rule is normalized $\overline{w_j}$ as the following equation (3.9)

$$\overline{w_j} = \frac{w_j}{\sum_{k=1}^{m} w_k} \tag{3.9}$$

This allows for the contribution of each rule to be proportional to each other in determining the final output.

d. Defuzzification Layer

Converts the fuzzy outputs of the rules into a crisp overall output. This is the combination of the output of the rules, obtained as a function of the inputs, and is often polynomial. It is expressed by the equation (3.10):

$$f_i(x) = \left(\sum_{i=1}^{n} p_{ij} x_i^2 + q_{ij} x_i + r_j\right) \tag{3.10}$$

Where p_{ij}, q_{ij}, and r_j are the parameters of the rule's output function.

e. Output Calculation Layer

It involves calculating the final output of the ANFIS model as an aggregate of the weighted outputs of each rule described by equation (3.11):

$$F(x) = \sum_{j=1}^{m} \overline{w_j}\, f_i(x) \tag{3.11}$$

Here, $f_i(x)$ denotes the output of the $j-th$ rule, and $\overline{w_j}$ its normalized firing strength.

3.5.2.2 Hybrid learning for parameter optimization

A hybrid learning strategy is utilized by the ANFIS model to optimize its parameters and improve the system's prediction potential.

a. Forward Pass: The forward pass involves examining the membership value for each input as the system goes through all the rules in a bid to check the strength of firing for all the rules defined. The different strengths need to be normalized. After normalization, the logical output level for every rule is obtained. This is the input value being used with the rule's specified consequent parameter.
b. Backward Pass (Gradient Descent Optimization): This phase of optimization includes the backward pass, which introduces parameter refinement, quantified through gradient descent. In turn, the backward pass intends to optimize premise parameters, which structure the membership functions, as well as the optimization of consequent parameters, which determine the contribution of rules' outputs. The backward pass is governed by the gradient of the loss function, which represents the difference between the model's predictions and the actual outcomes.
c. Parameter Update: Subsequently, the model goes through an update of its parameters so as to minimize the loss function. This process is vital in the determination of the most

effective way to update the parameters in order to accurately minimize the prediction error. The adjustments in the parameters are determined by the calculated gradients and the learning rate.

From the above comprehensive analysis, it is clear that the ANFIS framework efficiently and systematically optimizes the premise and consequent parameters. It rapidly enhances prediction accuracy by correcting consequent parameters during the forward pass and precisely adjusts premise parameters during the backward pass. Most model characteristics can, therefore, efficiently adapt to data trends, which ultimately improves predictive performance.

3.6 MODELS EVALUATION

In this research, it is crucial to determine the efficacy of the XGBoost and ANFIS machine learning models to clarify the extent of how they are reliable in predicting cardiovascular diseases. This step also allows measuring models' ability to predict new, unseen data, and therefore their applicability to clinical practice.

3.6.1 Performance Metrics

To quantify the models' predictive accuracy and reliability, a set of performance metrics will be employed, including [76]:

Confusion Matrix Analysis: An essential element of the evaluation is the use of a confusion matrix for each model, which is a comprehensive summary of the classification outcomes. Mathematically, the confusion matrix is structured as follows:

	$Predicted\ Positive$	$Predicted\ Negative$
$Actual\ Positive$	$True\ Positives\ (TP)$	$False\ Negatives\ (FN)$
$Actual\ Negative$	$False\ Positives\ (FP)$	$True\ Negatives\ (TN)$

The matrix allows for an all-inclusive examination of the classification model's performance by measuring the correct and incorrect predictions in relation to the positive and negative cases. This structured manner of analysis facilitates a more thorough insight into the model's predictive capabilities and shortcomings, that is, it measures the model's accuracy in recognizing and labelling the different instances within the dataset.

Accuracy: This metric is a direct indicator of the model's correctness in predicting outcomes, and it is calculated as the ratio of correctly predicted instances to the total instances, as the equation (3.12) below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.12}$$

Precision and Recall (Sensitivity): Precision is a statistical metric that calculates the ratio of correctly predicted positive outcomes to all predicted positive outcomes. It is represented by the equation (3.13):

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.13}$$

The recall metric evaluates the model's capacity to correctly identify all positive cases, as represented by equation (3.14):

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.14}$$

These metrics play a crucial role in medical diagnostics by striking a balance between the risk of missing genuine patients and the need to minimize false alarms.

F1 Score: The F1 Score is utilized as the harmonic measure of the precision and recall, and it provides a unified metric that considers both aspects into a single view of the model's success, allowing for increasing performance when a class representation imbalance. The F1 score is expressed as this equation (3.15):

$$F1\ Score = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{3.15}$$

F2 Score: which values recall higher than precision, is a modified metric used when false negative minimization is critical. It is a relevant metric to medical diagnostics; here, the cost of missing actual positive cases is higher than that of returning false positive results. It is computed as following equation (3.15):

$$F2\ Score = \frac{5*\text{Precision}*\text{Recall}}{4*\text{Precision}+\text{Recall}} \tag{3.16}$$

This equation modifies the F1 score to give additional weight to recall, ensuring that the model's ability to correctly identify all positive cases is prioritized in the evaluation process.

ROC-AUC Score: The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve provide insights into the model's discrimination capability, i.e., its ability to

distinguish between classes. A higher AUC indicates better model performance. It is which is calculated as the following equation (3.17):

$$\text{FPR} = \frac{FP}{FP+TN} \tag{3.17}$$

Feature Importance Analysis: Before applying the XGBoost and ANFIS model, feature importance analysis was carried out. It helps to determine the factors that have the most influence in predicting cardiovascular diseases. The feature importance analysis provides insights into the model's decision-making process and helps to determine the relative importance of features in the target variable.

Comparative Analysis: Both XGBoost and ANFIS models will be evaluated in the same manner, thus affording the possibility of comparative analysis. This evaluation will assist in identifying which of the models, would deliver the most precise and trustworthy predictions related to cardiovascular disease.

To sum up, the evaluation of XGBoost and ANFIS models using a thorough performance metrics set, such as the confusion matrix, accuracy, precision, recall, F1 score, and the ROC-AUC score, plays a significant role in understanding models' predictive worthiness in addressing cardiovascular diseases. Therefore, it is possible to assume that the thorough analysis not only provides an individual with the detailed results of predictive power of each model but also elucidates the practical potential of these models in real clinical settings. In the end, by investigating XGBoost and ANFIS models' performance in the cardiovascular diseases' prediction, the current work strives to establish which machine learning technique can best improve the accuracy and reliability of such predictions. In this way, the conclusions drawn from this analysis can help with the further development of ML tools for medicine, thereby improving patients' care strategies.

4. RESULTS AND ANALYSIS OF XGBOOST AND ANFIS MODEL

4.1 OVERVIEW

In this chapter, the results of applying the XGBoost and ANFIS models for cardiovascular disease prediction based on five different datasets, as well as analytical insights, are presented. The main objective of the chapter is to assess and compare the performance of the models to understand how these models can be effectively applied to clinical practice. Overall, the application of detailed analysis into the training process, parameter tuning, and performance establishing allows presenting the comprehensive picture of how XGBoost and ANFIS perform using multiple datasets for cardiovascular disease prediction. The results contribute to the understanding of the possibility to implement these models in combination with CVD prediction, presenting the critical overview of its strengths and limitations in the context of medical analytics.

4.2 XGBOOST MODEL ANALYSIS

4.2.1 Training and Testing Process for XGBoost

The training and testing regimen of the XGBoost model is instrumental in demonstrating its efficacy in predicting cardiovascular diseases across diverse datasets. These datasets, sourced from Switzerland, Cleveland, Hungarian, Long Beach VA, and Statlog, underscore the model's versatility and accuracy within varied healthcare data environments.

During the data preprocessing stage, categorical variables were encoded using OneHotEncoder, and numerical features were standardized with StandardScaler, thereby levelling the predictive landscape. To contend with missing values, KNNImputer was employed, striking a balance to avoid over- or underfitting our imputation approach.

Initial analysis of class distribution revealed significant imbalances, as the following Figure 4.1 for each dataset below.

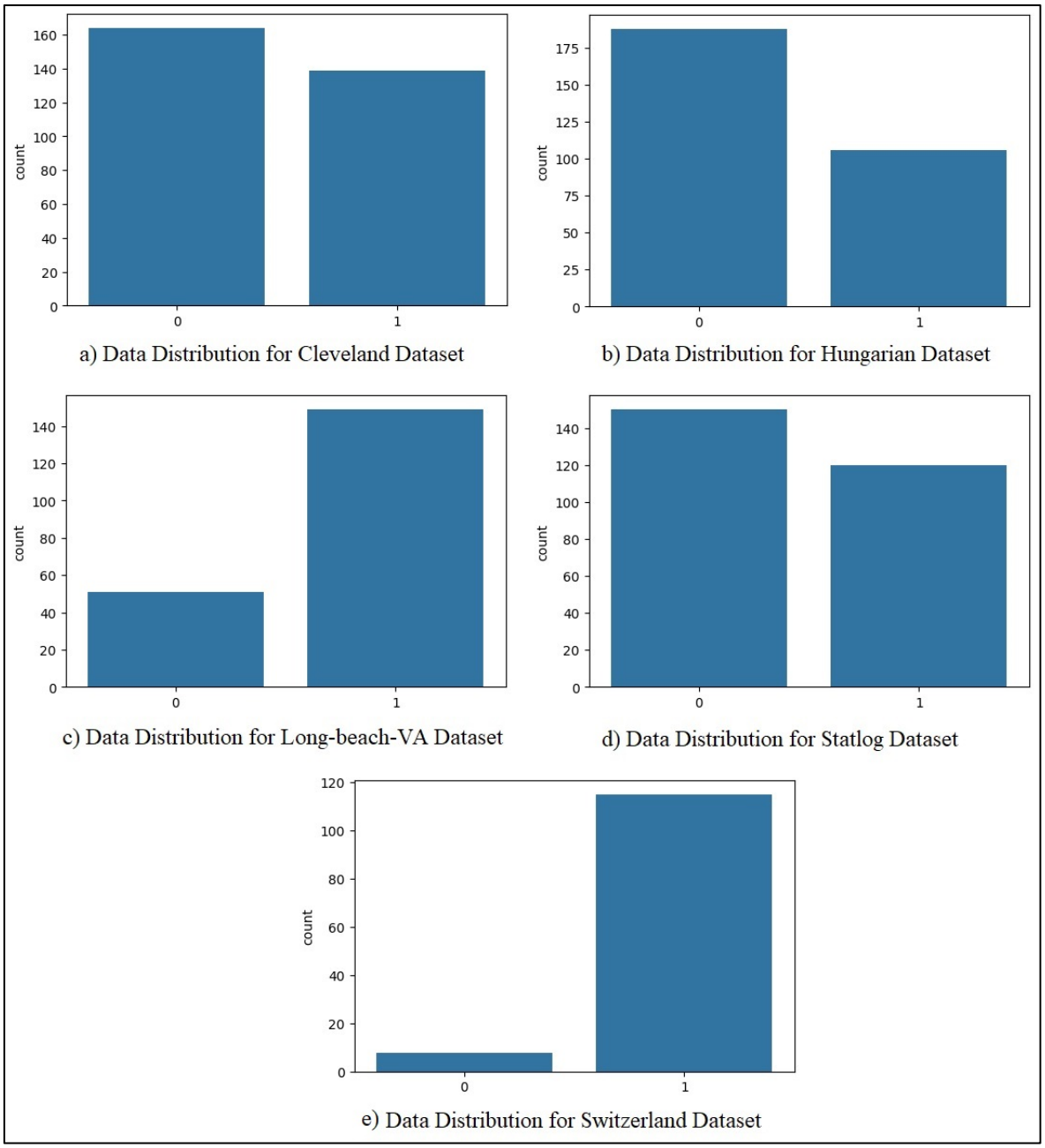

Figure 4.1: Original Data Distribution for Datasets.

To rectify the imbalance, the SMOTEENN technique was deployed, adeptly synthesizing data for the minority class and pruning the majority. This crucial step toward equality is reflected in the class frequencies' balanced distributions, as depicted in the subsequent Figure 4.2.

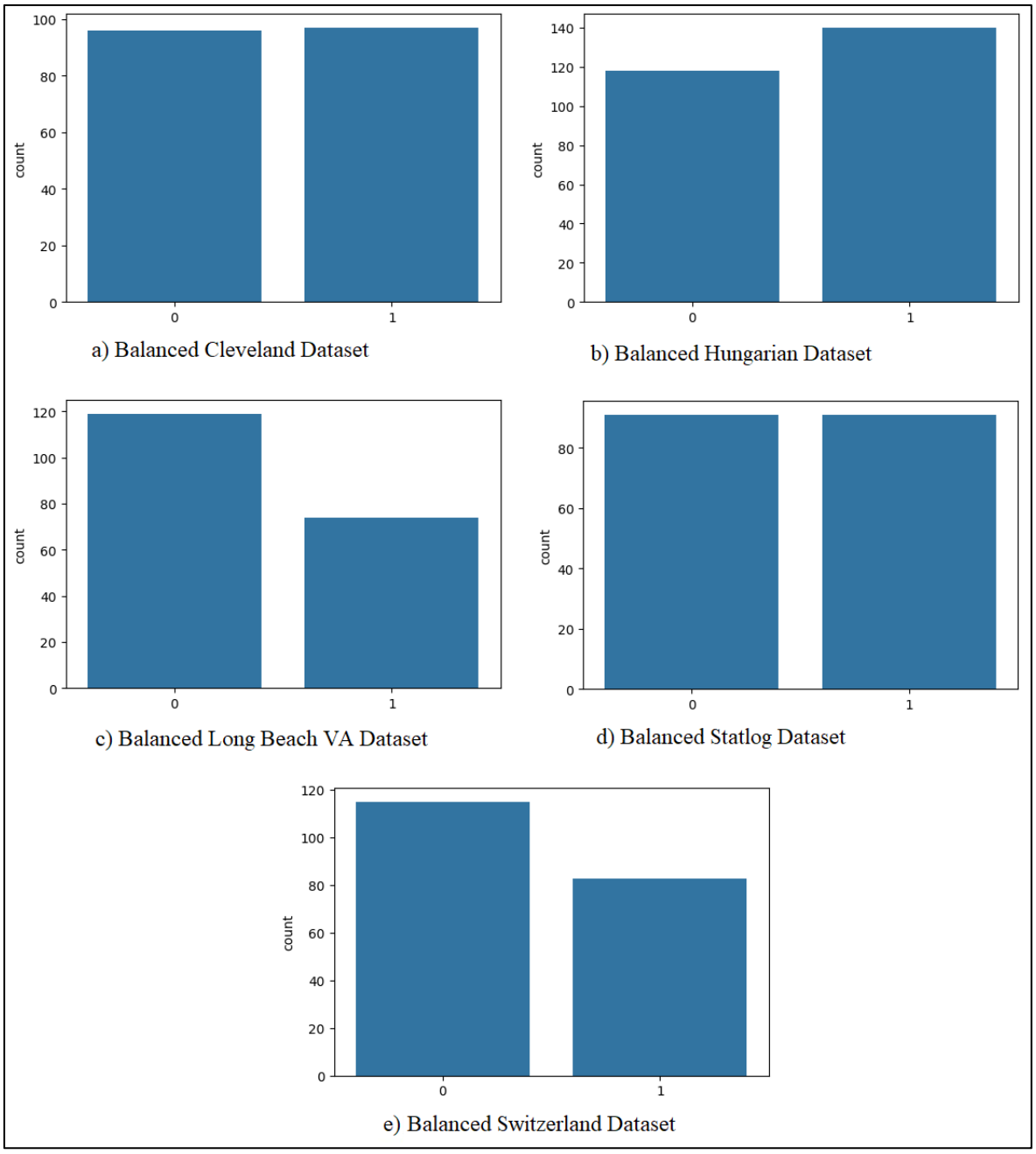

Figure 4.2: Data Distribution After SMOTEENN for Datasets.

Next, PCA was applied to the datasets to reduce dimensionality while maintaining data variance, which is vital for computational efficiency and model performance. The clustering revealed through the scatter plots in the figures (Figure 4.3,Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7) below provided insight into the data's underlying structure and class differentiation:

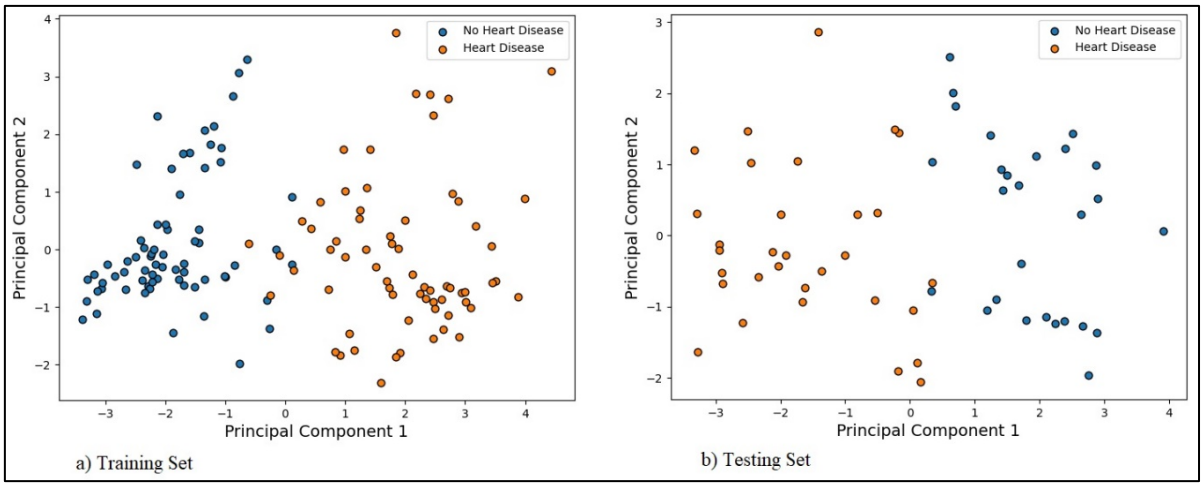

Figure 4.3: PCA Results for Cleveland Dataset.

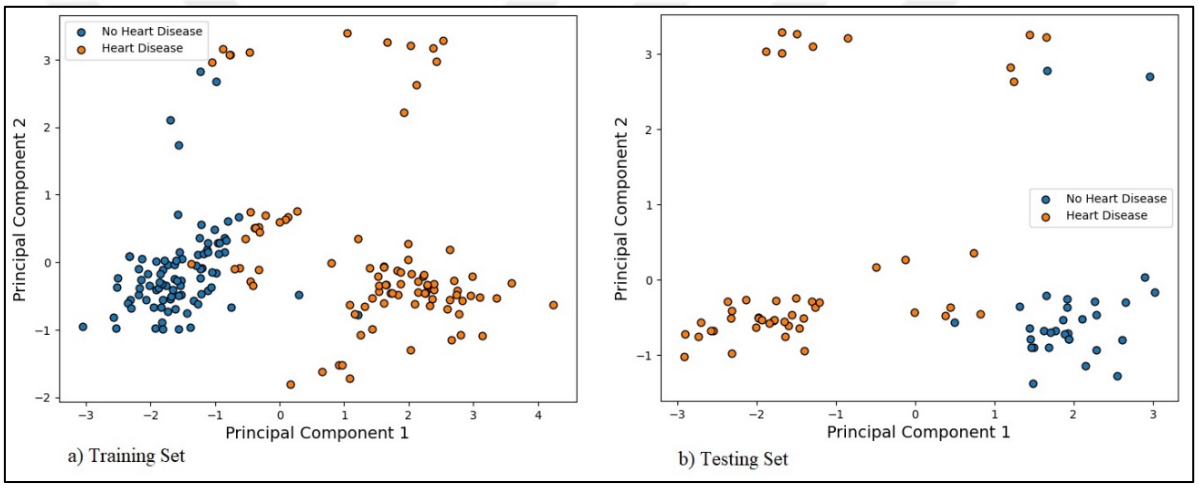

Figure 4.4: PCA Results for Hungarian Dataset.

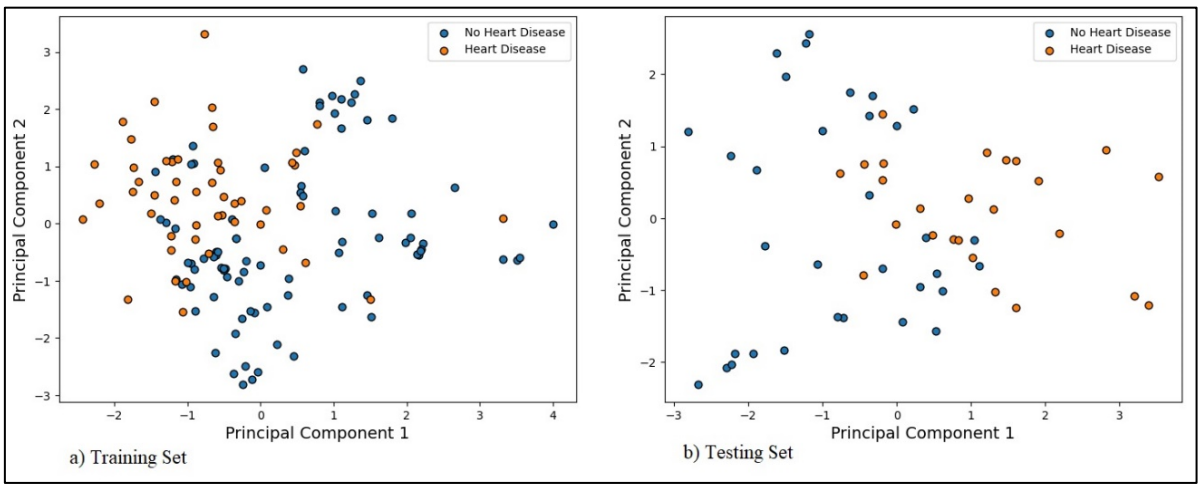

Figure 4.5: PCA Results for Long Beach VA Dataset.

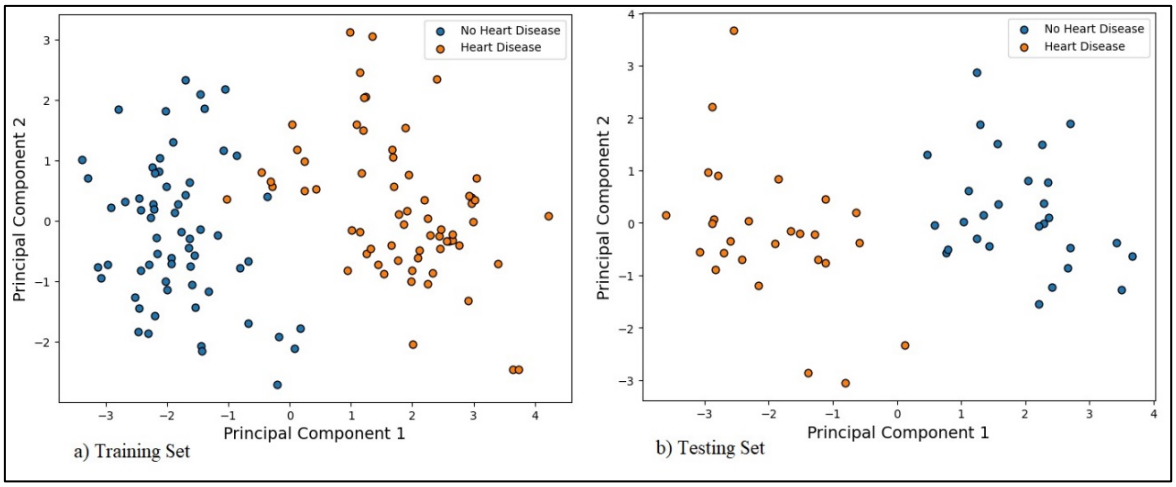

Figure 4.6: PCA Results for Statlog Dataset.

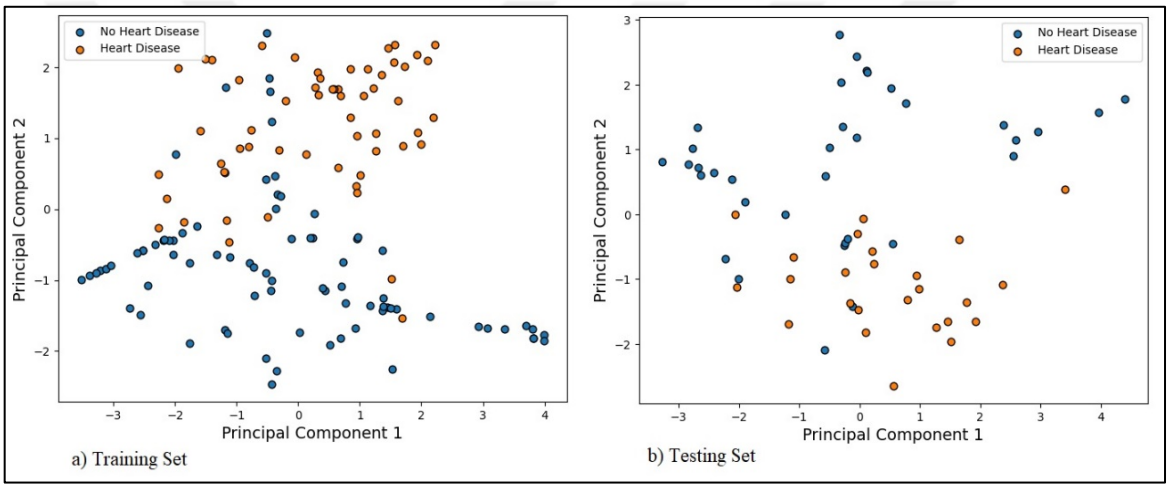

Figure 4.7: PCA Results for Switzerland Dataset.

These figures show how data points, now represented within a PCA-transformed space, and plotted along the principal components, suggest the model's potential for generalizing to unseen data.

For model training, a 70:30 split was chosen, with the larger portion for training the XGBoost model—selected for its robustness with varied data—and the lesser for testing. GridSearchCV's exhaustive search capabilities were leveraged to optimize the hyperparameters, targeting an improved F1 score for a balanced precision-recall trade-off. Following training, the model's threshold was fine-tuned using the F2 score to enhance sensitivity, a vital feature in medical diagnostics where the cost of missing a true positive is highly consequential.

4.2.2 Feature Importance with SelectFromModel

In the XGBoost model's predictive analysis for cardiovascular diseases, assessing the influence of individual features is essential. Our datasets included 14 distinct attributes as illustrate in the following Table 4.1:

Table 4.1: Clinical Attributes of Cardiovascular Datasets.

Attribute number	Attribute	Description
1	Age	Age of the patient
2	Sex	Biological sex of the patient
3	Chest pain type (cp)	Type of chest pain experienced
4	Resting blood pressure (trestbps)	Blood pressure in mm Hg on admission
5	Cholesterol (chol)	Serum cholesterol in mg/dl
6	Fasting blood sugar (fbs)	Blood sugar > 120 mg/dl (fasting)
7	Resting electrocardiogram (restecg)	Results of ECG at rest
8	Maximum heart rate (thalach)	Maximum heart rate achieved
9	Exercise-induced angina (exang)	Angina induced by exercise
10	ST depression (oldpeak)	ST depression induced by exercise
11	Slope of the peak exercise ST segment (slope)	Slope of the peak exercise ST segment
12	Number of major vessels (ca)	Number of major blood vessels stained
13	Thalassemia (thal)	Type of thalassemia

Utilizing SelectFromModel for feature selection, the model discerned which of these features were most predictive. Feature importance charts, both in visual and analytical form, shed light on the significance of these variables in the model’s assessment of cardiovascular risk.

Cleveland Dataset Feature Importance: For the Cleveland dataset, the attribute of age was a standout factor, indicating its substantial role in cardiovascular health assessments for these patients. The corresponding feature importance graph Figure 4.8 displays this attribute's significant impact on the model's predictive accuracy.

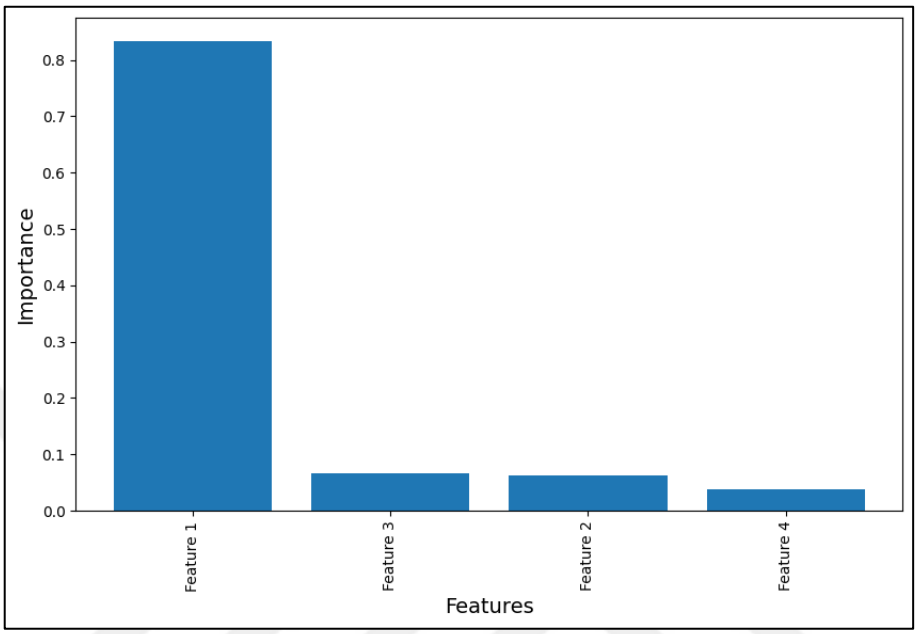

Figure 4.8: Feature Importance for Cleveland Dataset.

Hungarian Dataset Feature Importance: The Hungarian dataset's feature importance spanned across several clinical attributes, signaling a multifaceted set of factors like chest pain type and blood sugar levels that collectively influence heart disease risks, as illustrated in Figure 4.9.

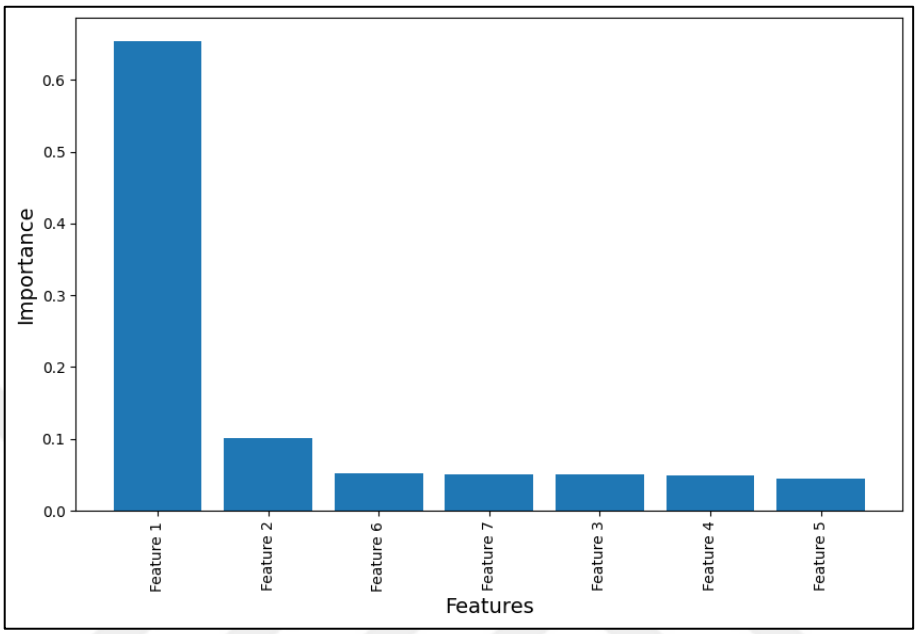

Figure 4.9: Feature Importance for Hungarian Dataset.

Long Beach VA Dataset Feature Importance: In the Long Beach VA dataset, the peak heart rate achieved during exercise was pivotal, suggesting particular cardiovascular concerns for this dataset, which are highlighted in Figure 4.10.

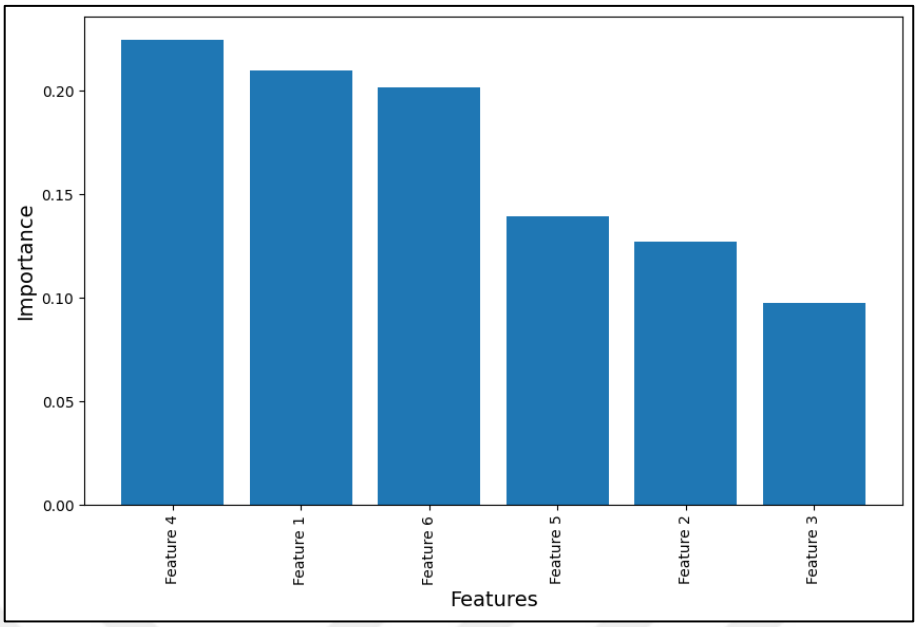

Figure 4.10: Feature Importance for Long Beach VA Dataset.

Statlog Dataset Feature Importance: The Statlog dataset placed emphasis on genetic factors such as thalassemia and anatomical considerations like the number of vessels, as seen in Figure 4.11, emphasizing their importance in heart disease prediction.

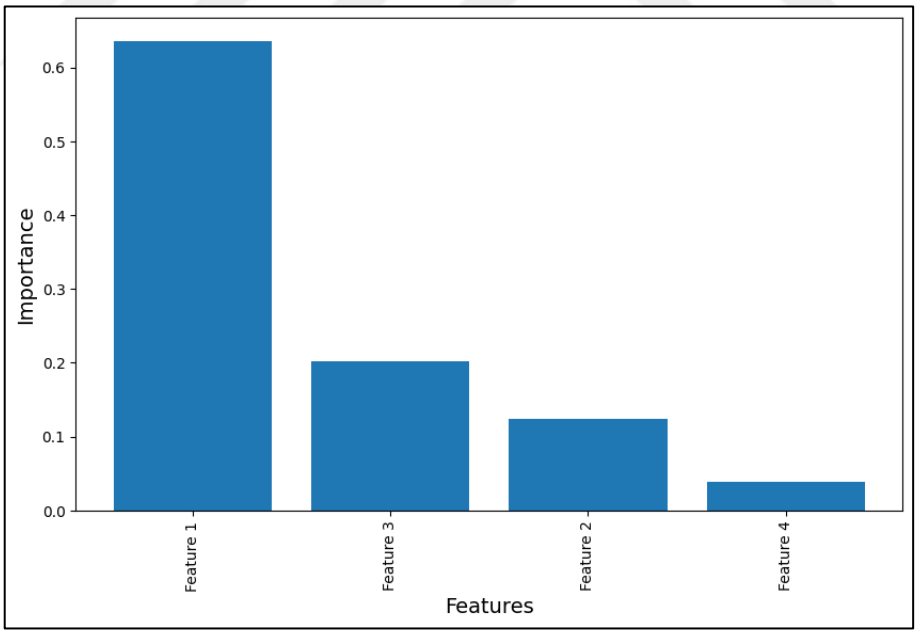

Figure 4.11: Feature Importance for Statlog Dataset.

Switzerland Dataset Feature Importance: Attributes related to the patients' response to exercise, specifically angina and ST depression, were significant in the Switzerland dataset's predictive modeling, as shown in Figure 4.12.

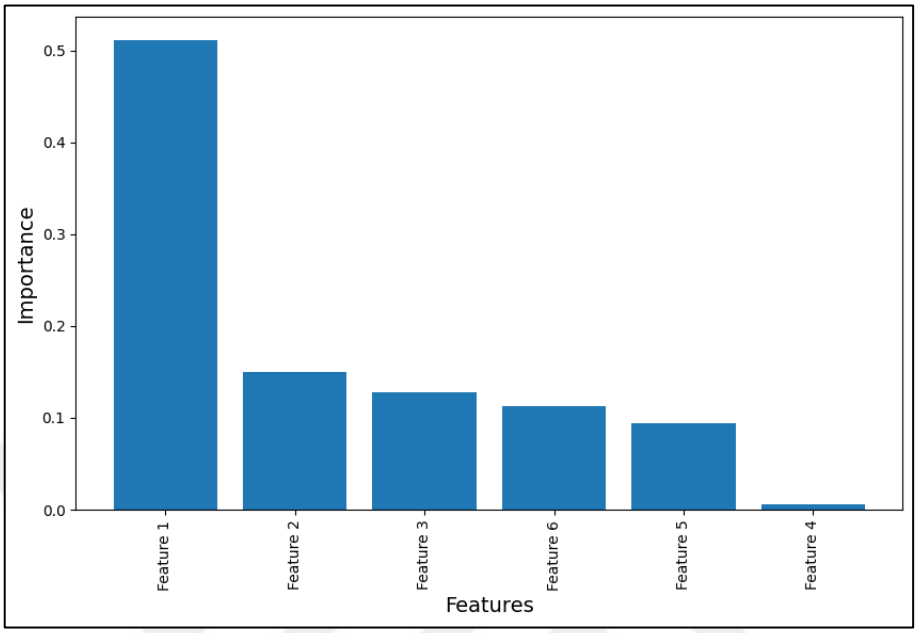

Figure 4.12: Feature Importance for Switzerland Dataset.

The visualized data underscores not just the variables most predictive of cardiovascular disease but also the intricate relationship between patient traits and the onset of heart conditions. By correlating the XGBoost model's outputs with specific clinical meanings, health professionals gain a targeted approach to addressing the most influential factors, paving the way for more precise and impactful healthcare interventions.

4.2.3 Performance Metrics and Results

In this section, we explore the performance metrics of the XGBoost algorithm, reflecting its accuracy and reliability in predicting cardiovascular diseases across five datasets. The metrics chosen accuracy, precision, recall, F1 score, F2 score, and ROC-AUC score serve as the benchmarks for evaluating the model's performance.

Cleveland Dataset Performance: Analyzing the Cleveland dataset, the XGBoost model's accuracy was measured at 98.28%. A precision of 0.97 and a perfect recall indicate the model’s reliable performance in identifying patients with and without the disease. The F1 score, which balances precision and recall, was calculated at 0.98. The F2 score, placing

more emphasis on recall, also showed a high value of 0.99. These scores collectively suggest that the model effectively prioritizes the correct identification of disease presence.

The ROC-AUC score of 0.98 confirms the model's discriminative capability. As depicted in Figure 4.13, the test ROC curve approaches the ideal with an area of 0.99, reflecting the model's precision in classifying test data.

Figure 4.14 presents the confusion matrix for the Cleveland dataset, where the model correctly predicted 26 instances as negative (true negatives) and 31 as positive (true positives), with a single case incorrectly predicted as positive (false positive). This matrix validates the calculated precision and recall, providing a tangible representation of the model's performance.

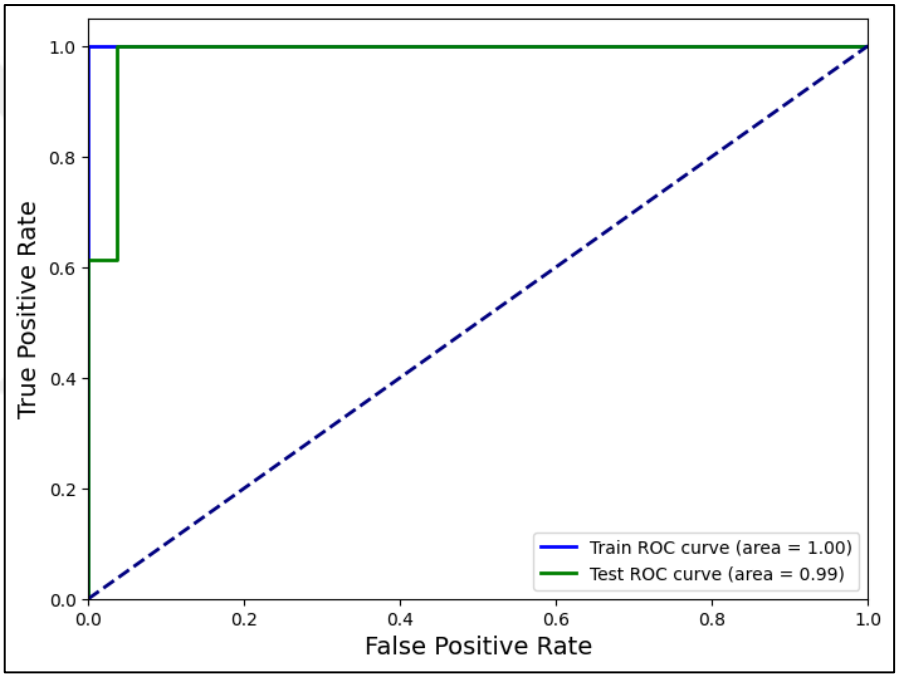

Figure 4.13: ROC Curve for Cleveland Dataset.

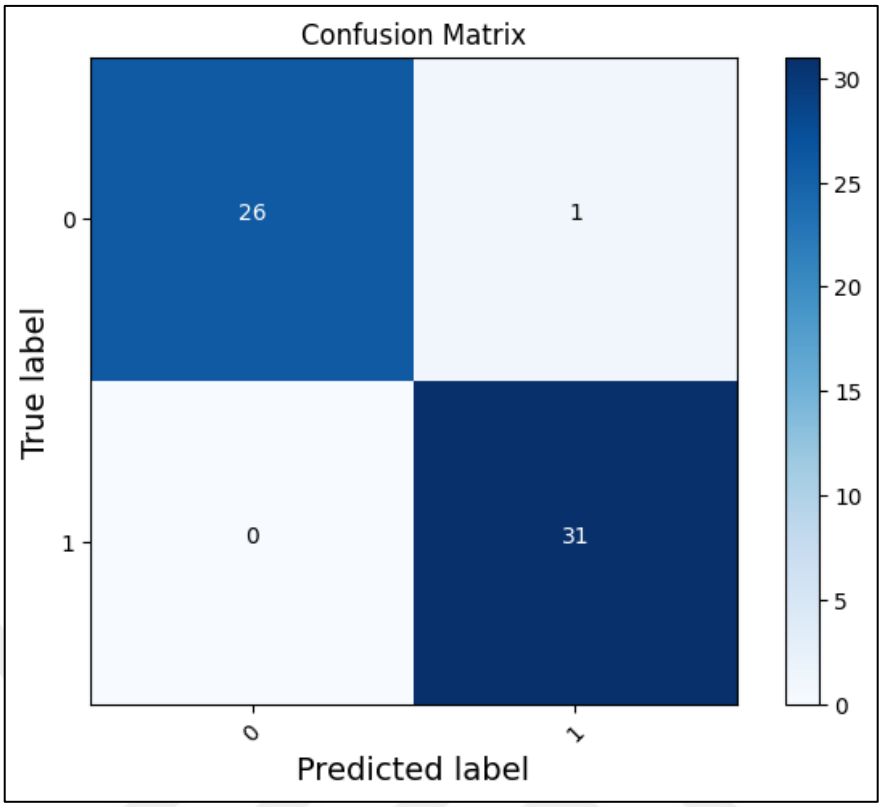

Figure 4.14: Confusion Matrix for Cleveland Dataset.

Hungarian Dataset Performance: The performance evaluation of the XGBoost model on the Hungarian dataset indicated a high level of accuracy at 99%. The precision rate stood at 0.98, and recall was calculated to be perfect, pointing to the model's strong predictive reliability. This led to an F1 score of 0.99, effectively capturing the balance between precision and recall, and the F2 score achieved a maximum value of 1.00, highlighting the model's sensitivity in identifying true positive cases.

The ROC-AUC score was also notable at 0.98, a figure that reflects the model's ability to accurately discriminate between the disease classes. This is visually represented in Figure 4.15, where both the train and test ROC curves display an area of 1.00, indicative of the model's precision in classification.

The confusion matrix, shown in Figure 4.16, corroborates these metrics by displaying a total of 30 true negatives and 47 true positives accurately identified by the model, with only one instance of a false positive. The absence of false negatives in this matrix is consistent with the perfect recall score, underscoring the model's capability in identifying all positive cases.

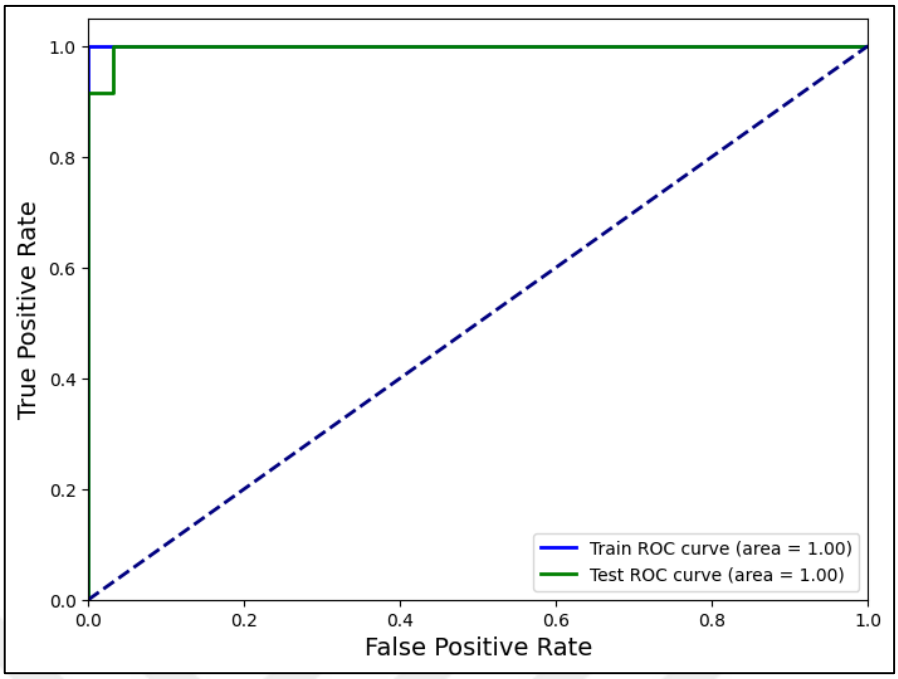

Figure 4.15: ROC Curve for Hungarian Dataset.

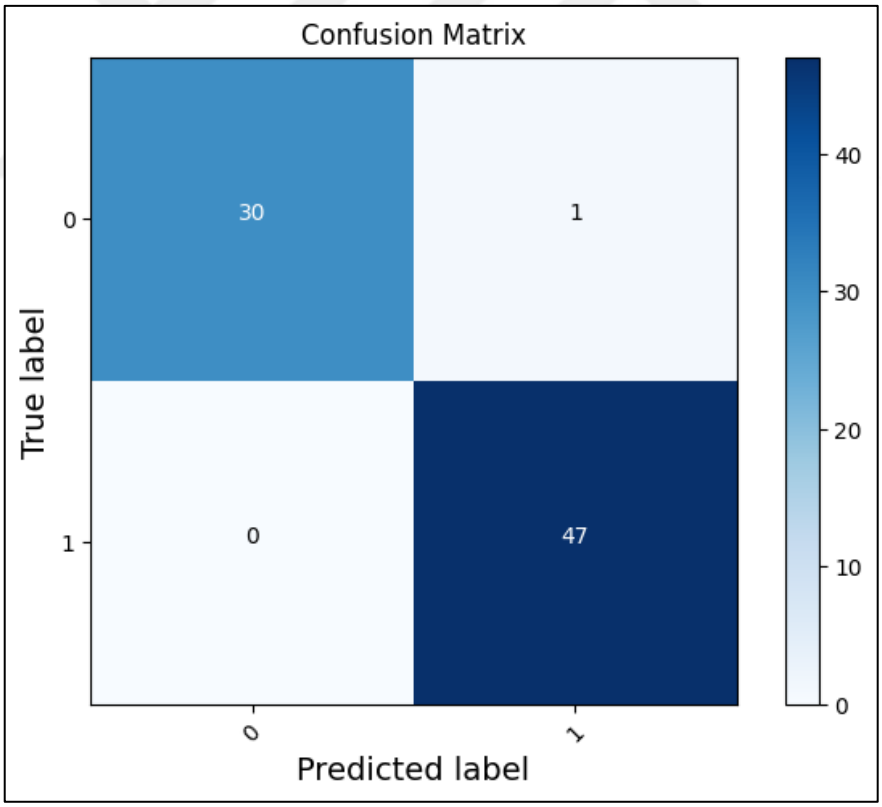

Figure 4.16: Confusion Matrix for Hungarian Dataset.

Long Beach VA Dataset Performance: On the Long Beach VA dataset, the XGBoost model demonstrated a reliable accuracy of 97%. The model was precise in its predictions with a precision score of 0.96 and equally robust in recall, accurately identifying those with the disease. The F1 score, at 0.96, indicates a balanced mean of precision and recall, while the F2 score, also at 0.96, reflects the model's consistent emphasis on correctly identifying positive cases.

The ROC-AUC score for this dataset was 0.96, suggesting the model's effectiveness at distinguishing between positive and negative cases. The ROC curve, shown in Figure 4.17, demonstrates this capability, with the area under the test ROC curve nearly matching that of the training, reinforcing the model's consistency.

The confusion matrix, depicted in Figure 4.18, validates the model's accuracy and reliability, showing 32 true negatives and 24 true positives. A single instance was incorrectly predicted in each of the positive and negative categories, indicating the model's high precision and recall in practical terms.

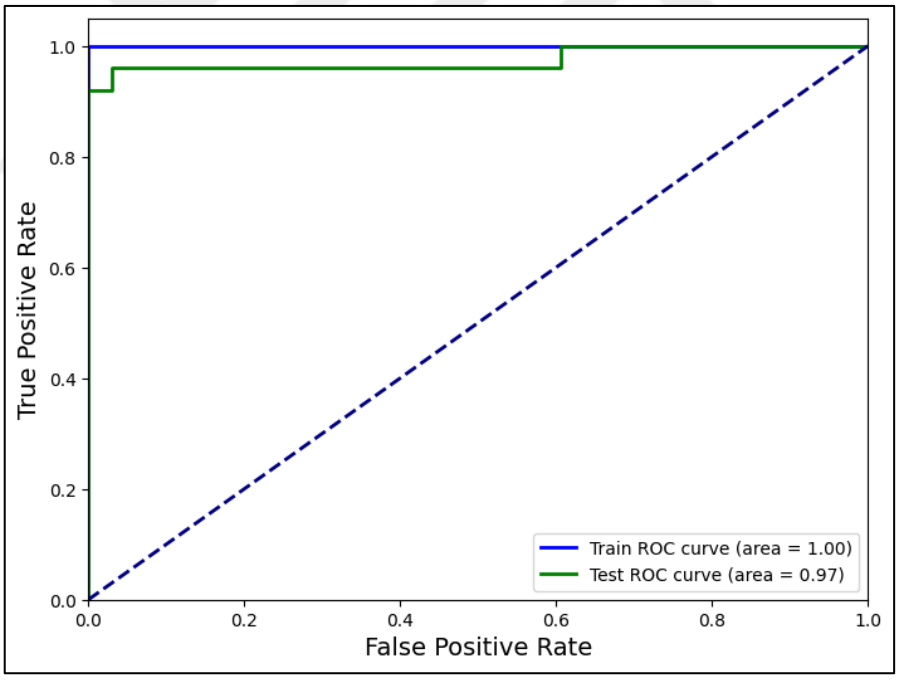

Figure 4.17: ROC Curve for Long Beach VA Dataset.

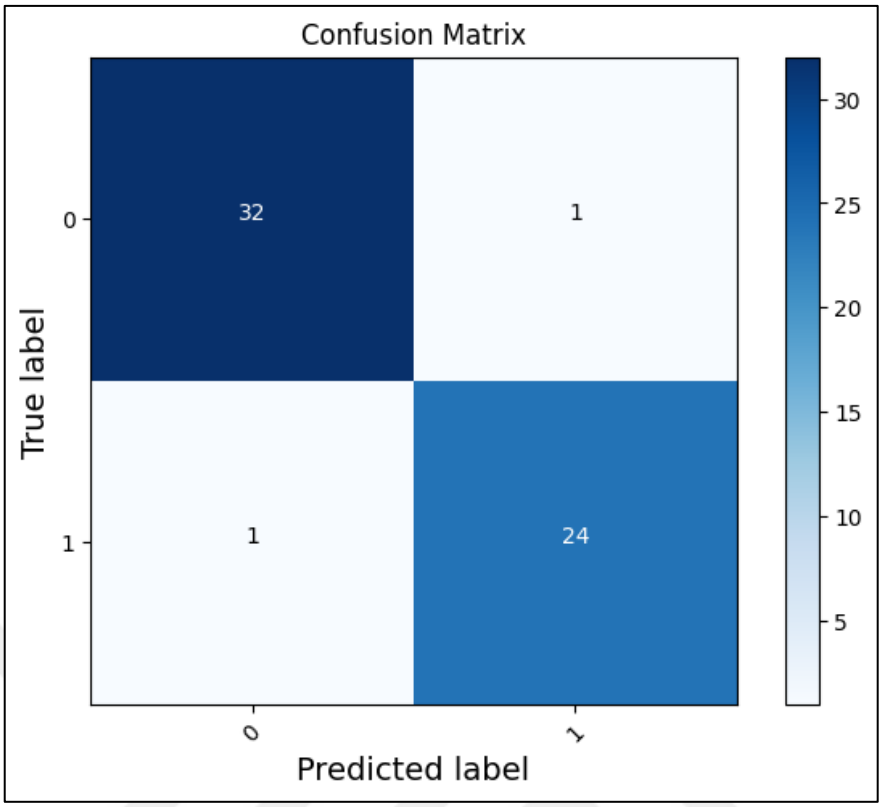

Figure 4.18: Confusion Matrix for Long Beach VA Dataset.

Switzerland Dataset Performance: Evaluating the Switzerland dataset, the XGBoost model delivered optimal performance across all fronts. It achieved a flawless accuracy of 100.00%, where precision and recall both reached the maximum possible score, indicating no misclassifications were made. This precision is reflected in the F1 and F2 scores, which both attained a perfect score of 1.00, denoting exceptional model accuracy in both the identification of true positives and the correct rejection of negatives.

The model's ROC-AUC score, which measures its ability to discriminate between the classes, was also perfect at 1.00. The corresponding ROC curve, displayed in Figure 4.19, further validates this with both the training and test curves achieving the maximum area under the curve, demonstrating that the model's predictions were consistently accurate.

In Figure 4.20, the confusion matrix for the Switzerland dataset paints a clear picture of the model's performance. It correctly predicted all cases without any false positives or false negatives, as indicated by the counts of 35 true negatives and 25 true positives. This level of accuracy showcases the model's capability in this specific dataset and underscores its potential utility in a clinical setting.

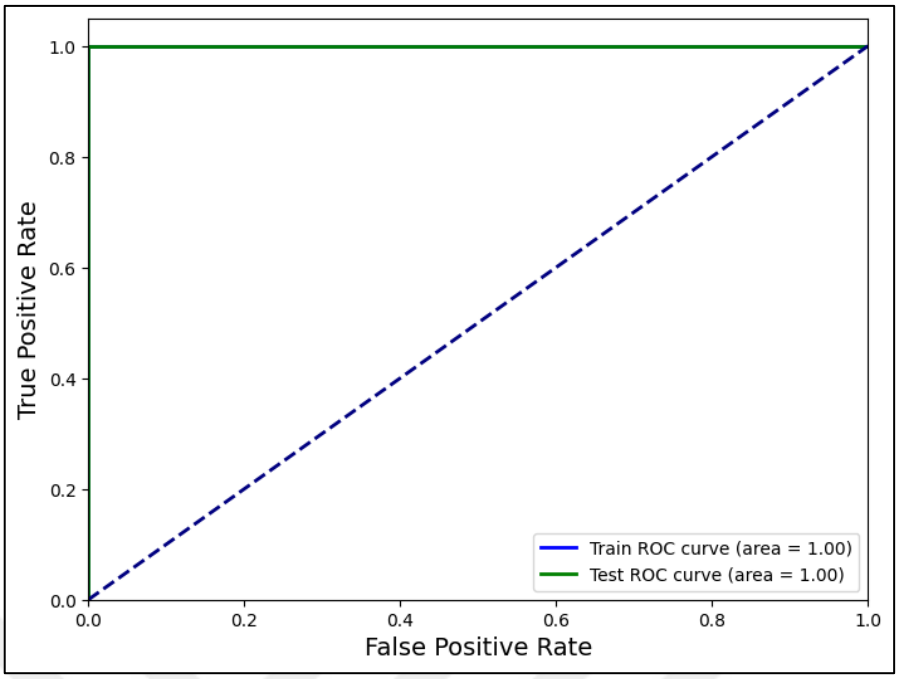

Figure 4.19: ROC Curve for Switzerland Dataset.

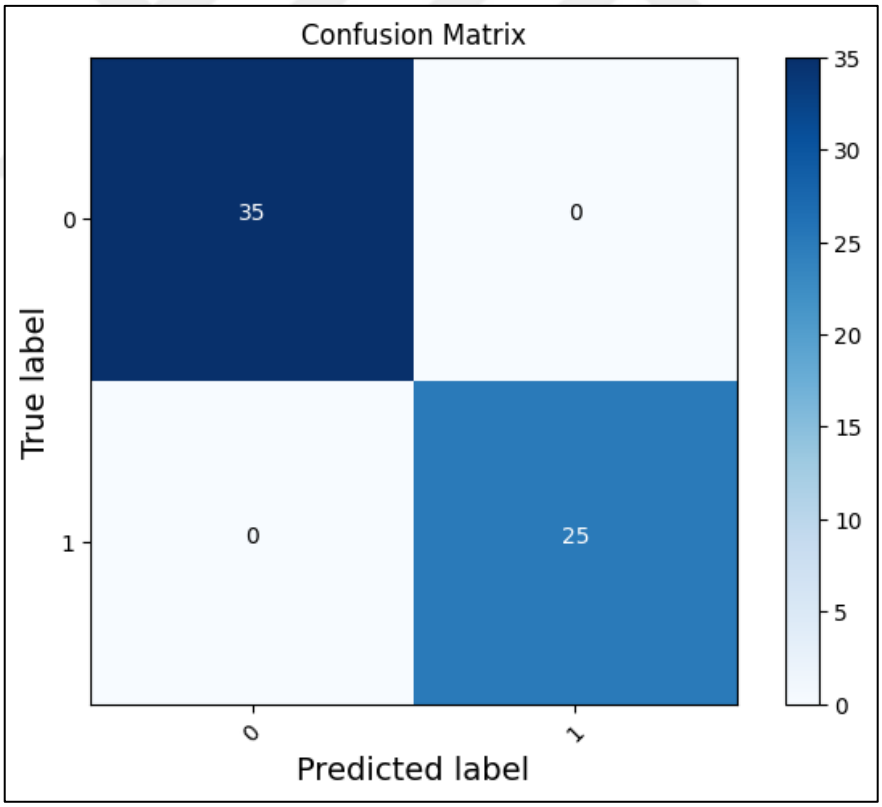

Figure 4.20: Confusion Matrix for Switzerland Dataset.

Statlog Heart Dataset Performance: On the Statlog Heart dataset, the XGBoost model achieved a perfect record of prediction accuracy, with both accuracy and precision rates

hitting the 100.00% mark. The model's ability to identify all positive cases was flawless, as indicated by a recall score of 1.00. This level of precision is echoed in both the F1 score and the F2 score, which take into account precision and recall, confirming the model's absolute accuracy in predicting cardiovascular disease outcomes for this dataset.

The ROC-AUC score maintained this trend of excellence, reaching 1.00, which signifies the model's exceptional ability to discriminate between positive and negative cases with complete accuracy. Figure 4.21 presents the ROC curve for the Statlog dataset, where the area under both the train and test curves hits the ideal mark, reinforcing the model's consistency in performance.

Complementing the ROC analysis, Figure 4.22 displays the confusion matrix for the Statlog Heart dataset. This matrix shows that the model precisely predicted 28 true negatives and 27 true positives, with no instances of false positives or false negatives. The clean division in this confusion matrix underscores the model's pinpoint precision in classification tasks.

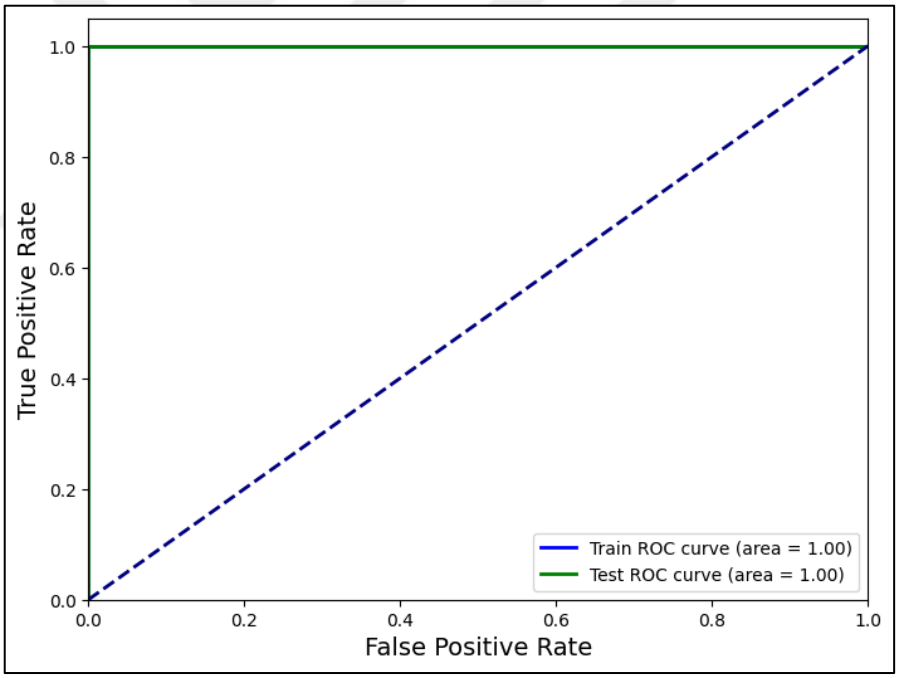

Figure 4.21: ROC Curve for Statlog Dataset.

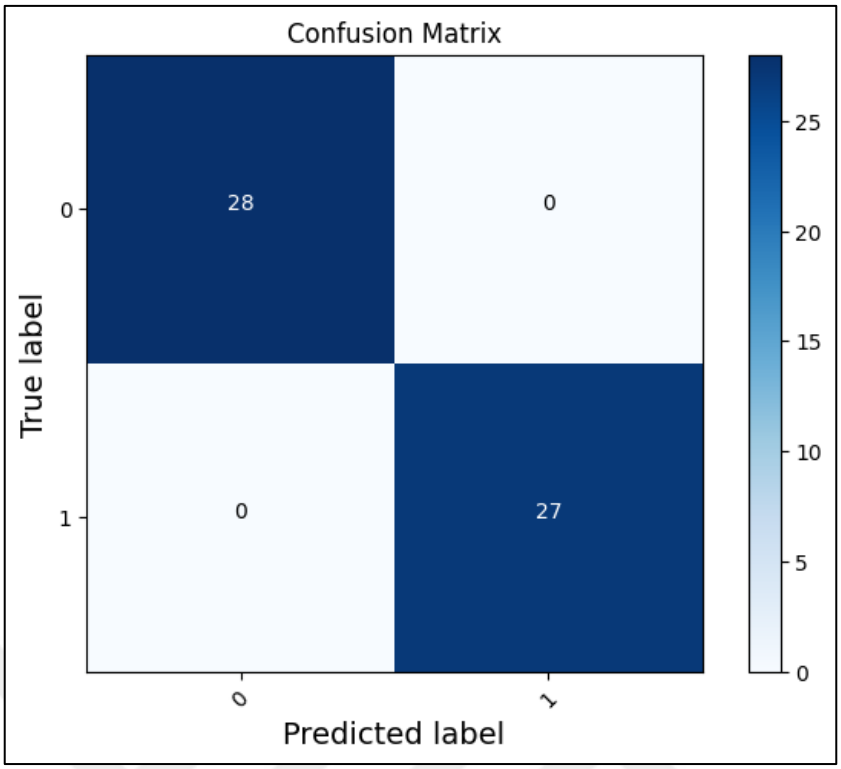

Figure 4.22: Confusion Matrix for Statlog Dataset.

The outcomes from the XGBoost model assessments across diverse datasets demonstrate its exemplary performance in identifying cardiovascular diseases with remarkable accuracy. The model has consistently maintained high precision and recall, ensuring a dependable balance in its predictive capabilities. Such uniformity is evidenced by the uniformly high ROC-AUC scores, emphasizing the model's adeptness at differentiating between patients with and without the condition. These steadfast metrics underscore the model's reliability and its suitability as a predictive instrument in clinical environments for the prognosis and diagnosis of cardiovascular ailments.

4.3 ANFIS MODEL ANALYSIS

The ANFIS model stands out for its integration of neural networks with fuzzy logic, offering a powerful approach to deciphering complex interdependencies within medical data. The analysis using ANFIS across five distinct cardiovascular datasets showcases its potential to address challenging diagnostic problems.

4.3.1 Training and Testing Overview

The ANFIS model's journey through the training and testing phases was rigorous, aiming to refine its proficiency for cardiovascular disease diagnosis. Initially, each dataset underwent a comprehensive preprocessing routine, paralleling the methodology applied to the XGBoost analysis. This standardization was crucial to ensure comparability of results.

The datasets initially presented imbalances in class distribution, which could potentially skew the model's learning. To address this, the SMOTEENN technique was employed, enhancing the representation of minority classes, and removing any ambiguous instances. This preprocessing step ensured that the datasets were well-defined and balanced, as depicted by the distribution visualizations in Figure 4.23.

Once balanced, as illustrated in Figure 4.24, the datasets were ready for the model's learning phase. The ANFIS model engaged in an extensive training process, tuning its internal parameters, such as the membership function parameters and consequent coefficients over a series of epochs. A carefully chosen learning rate facilitated a steady yet comprehensive learning curve, avoiding the pitfalls of overfitting or premature convergence.

Derived directly from the data, the model's rules encapsulated the nuanced relationships within the datasets. The ANFIS framework's ability to emulate human-like reasoning through fuzzy logic was especially beneficial, given the complexity often found in medical data.

Throughout this phase, ANFIS demonstrated its capacity to learn and adapt, underscoring the potential of such models in capturing the subtleties of medical diagnosis and contributing valuable insights into cardiovascular disease prediction.

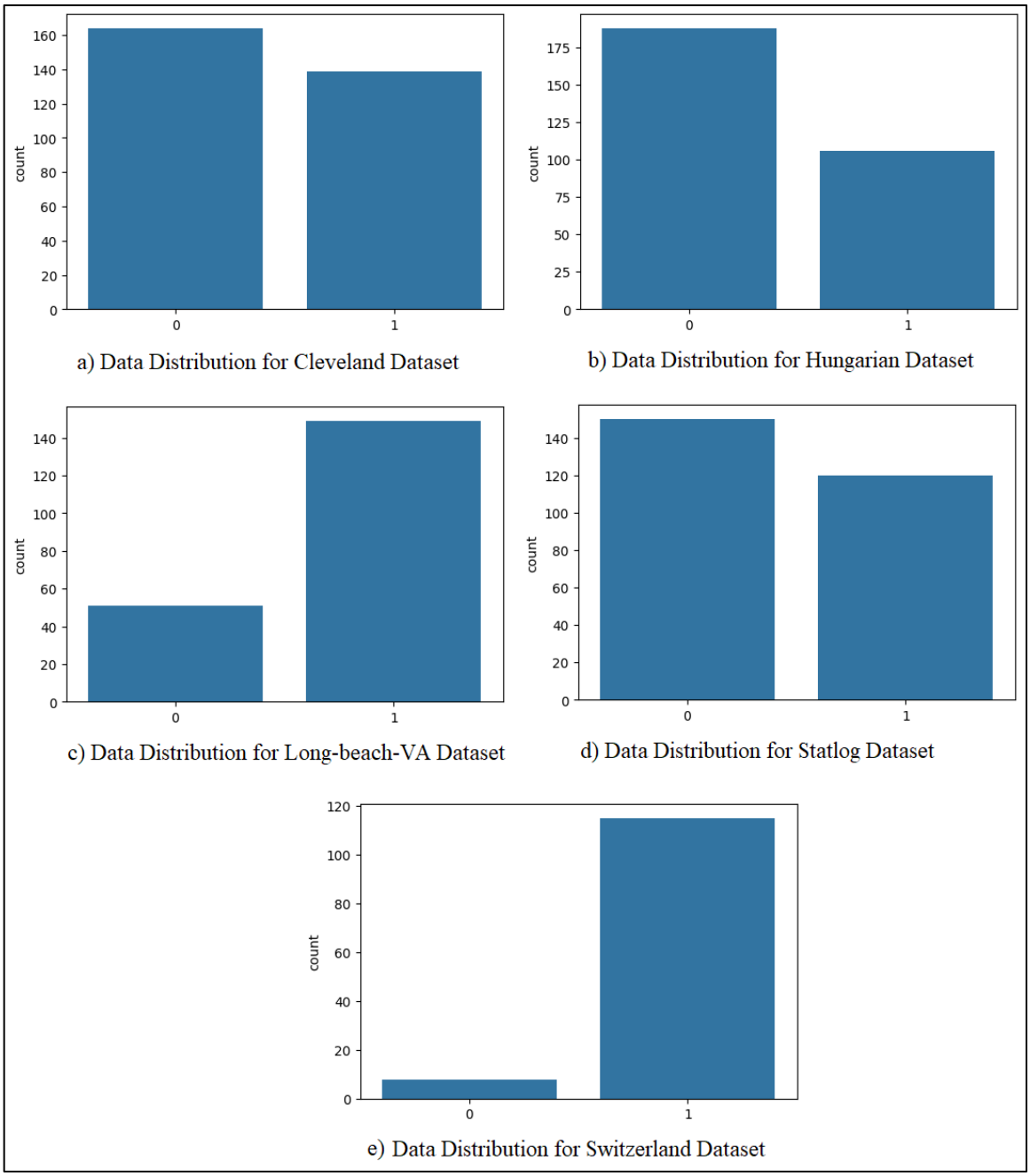

Figure 4.23: Original Data Distribution for Datasets in ANFIS.

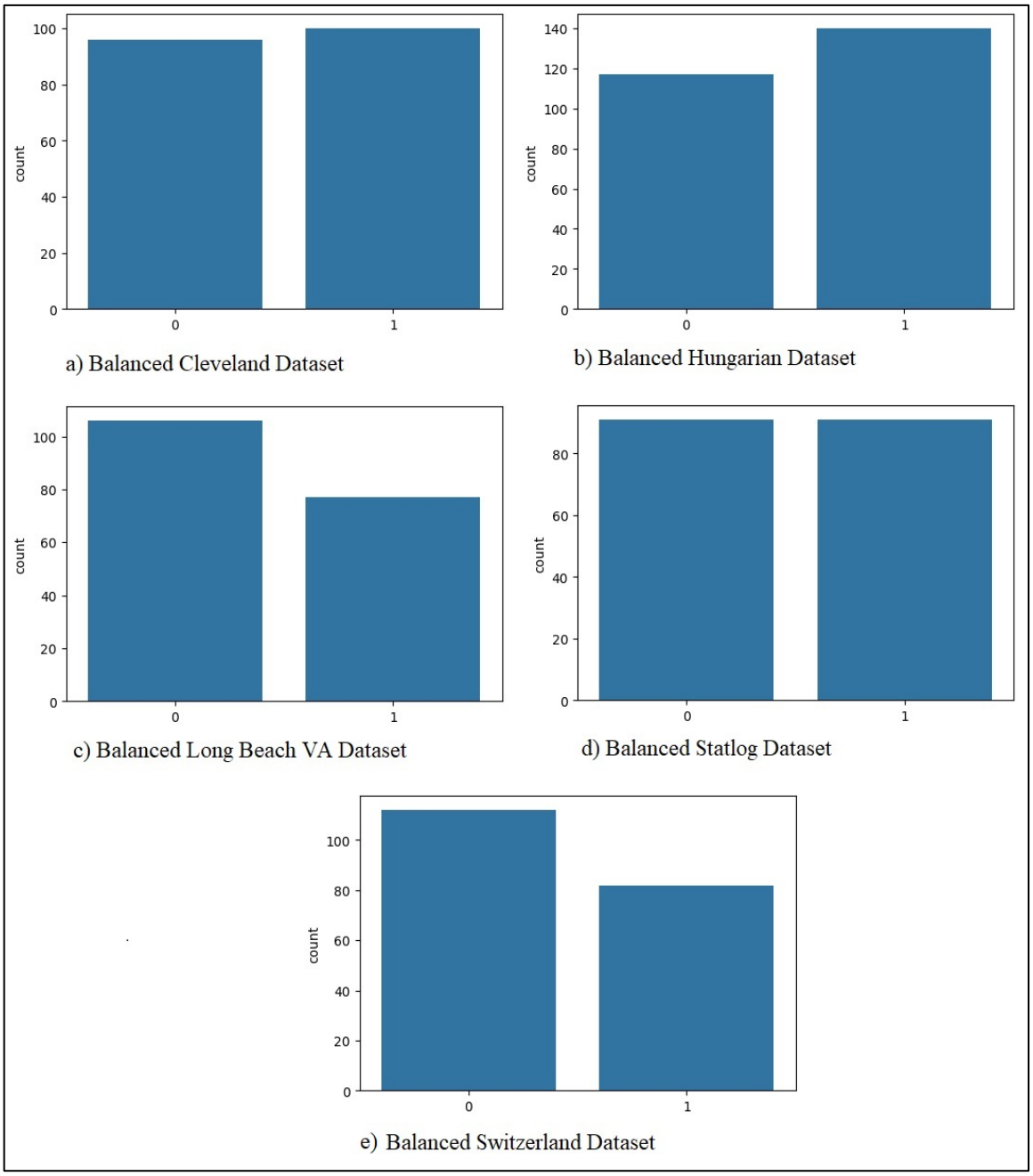

Figure 4.24: Data Distribution After SMOTEENN for Datasets in ANFIS.

With the data prepared, the ANFIS model was trained over numerous epochs. It learned and adjusted its parameters the centers and widths of the sigmoidal membership functions and the coefficients of the quadratic consequents to map the input features to the output classes with increasing accuracy. The learning rate was carefully selected to balance the speed of convergence with the model's ability to navigate the complex error landscape without becoming trapped in local minima.

The model's rules were derived from the data, allowing it to create a fuzzy logic-based framework that could capture the intricate, nonlinear relationships inherent in the diagnostic data. This setup provides a powerful method for medical diagnostic tasks, where the interplay of symptoms and test results is often complex and not well-suited to linear models.

4.3.2 Feature Importance with Recursive Feature Elimination

In the process of refining the ANFIS model for predicting cardiovascular disease, the technique of Recursive Feature Elimination (RFE) was employed to ascertain which specific clinical and demographic features were most instrumental in influencing the model's predictions across various datasets.

The RFE analysis on the Cleveland dataset illuminated that feature 13, representing thalassemia (thal), emerged as the most influential in the model's decision-making process, as evident in the corresponding graph Figure 4.25 Such an observation suggests thalassemia's significant role in cardiovascular risk within this patient cohort.

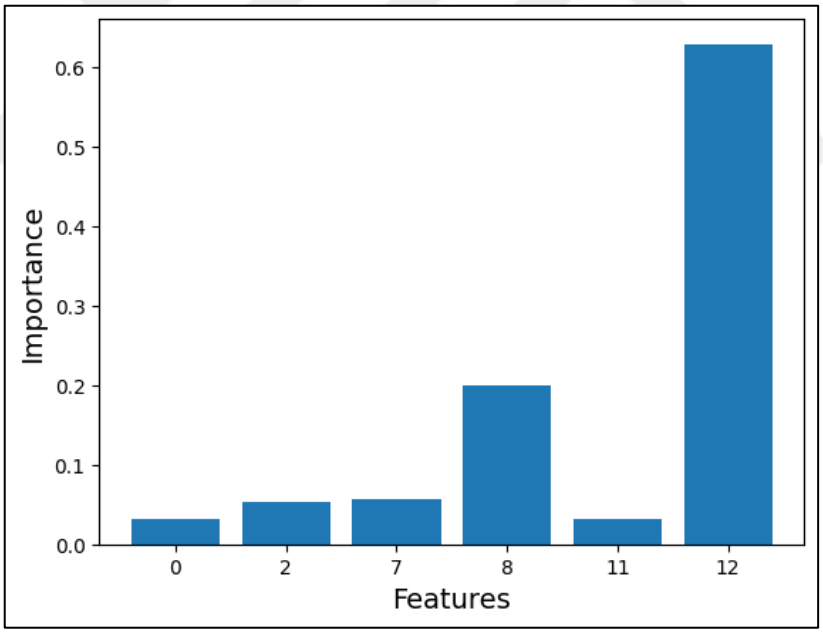

Figure 4.25: Feature Importance Analysis for Cleveland Dataset with RFE.

In the case of the Hungarian dataset, feature 11, which corresponds to the slope of the peak exercise ST segment (slope), was identified as the most impactful on the model's predictive power, showcased in the Feature Importance graph Figure 4.26. This underscores the clinical relevance of exercise-related ECG findings in assessing heart disease risk.

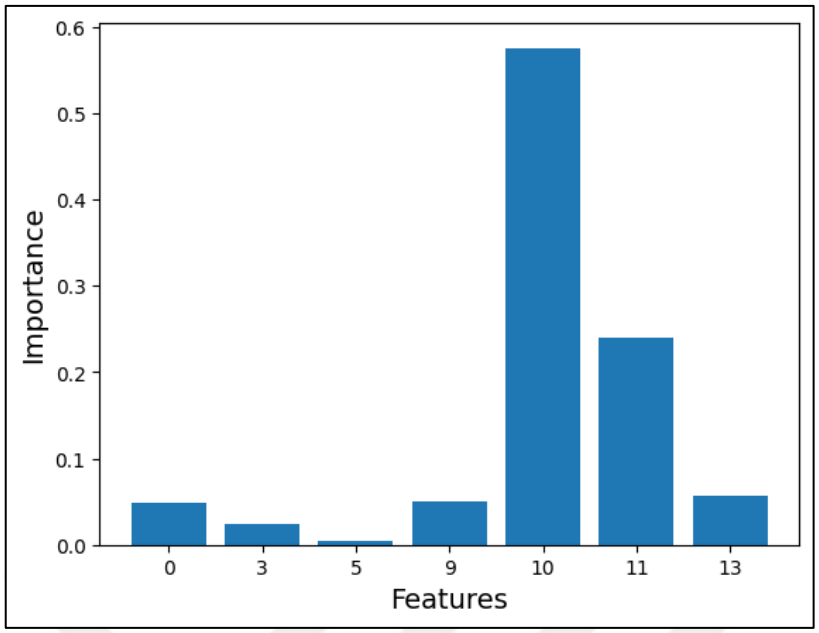

Figure 4.26: Feature Importance Analysis for Hungarian Dataset with RFE.

For the Long Beach VA dataset, it was feature 10, ST depression induced by exercise relative to rest (oldpeak), that stood out in the RFE findings, indicating its pivotal role in the model's accuracy for this demographic group, as depicted in Figure 4.27.

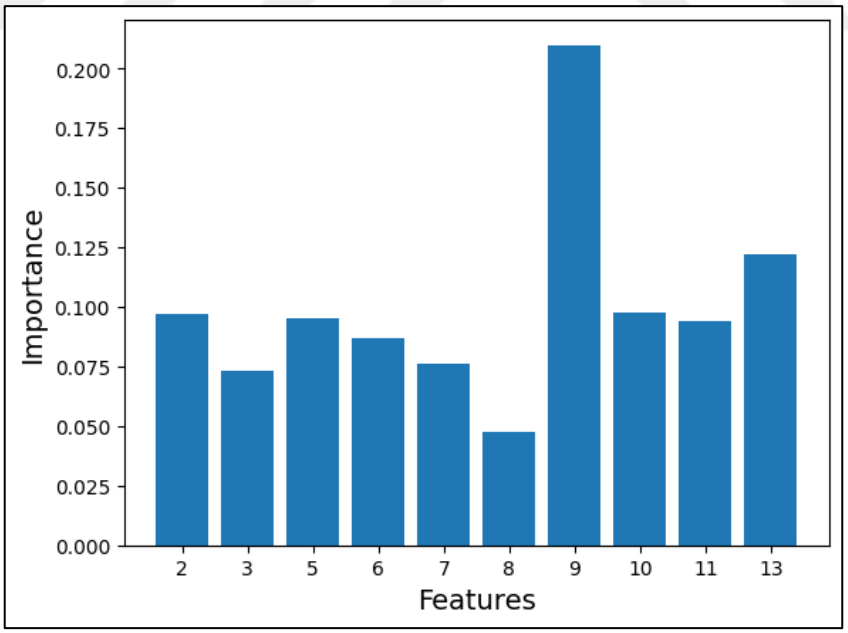

Figure 4.27: Feature Importance Analysis for Long Beach VA Dataset with RFE.

The analysis for the Statlog dataset revealed that feature 13, thalassemia (thal), was again a predominant factor, suggesting a consistency in the importance of this feature across different datasets, demonstrated in Figure 4.28.

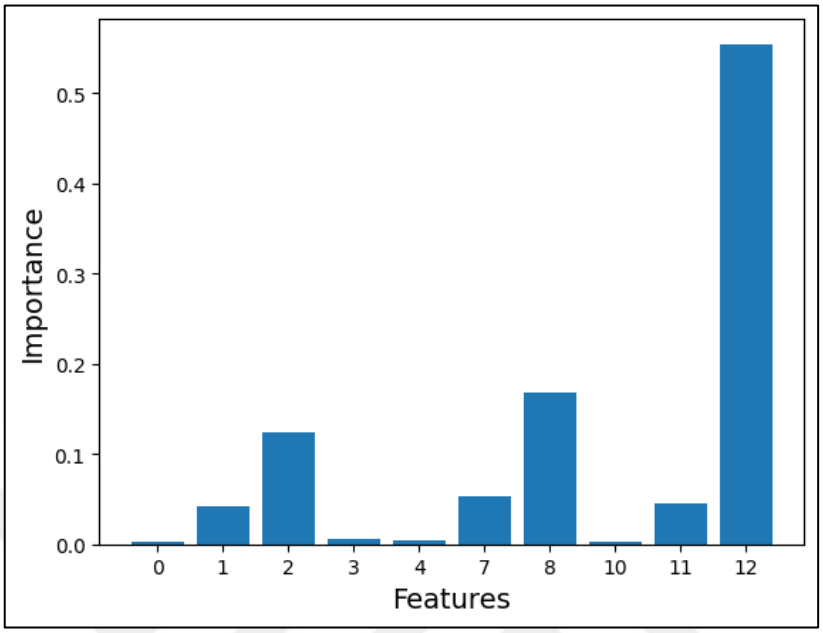

Figure 4.28: Feature Importance Analysis for Statlog Dataset with RFE.

Finally, the RFE results for the Switzerland dataset highlighted feature 4, resting blood pressure (trestbps), as a critical determinant in the model's predictions, signifying the impact of blood pressure on heart disease risk in this population group, as illustrated in Figure 4.29.

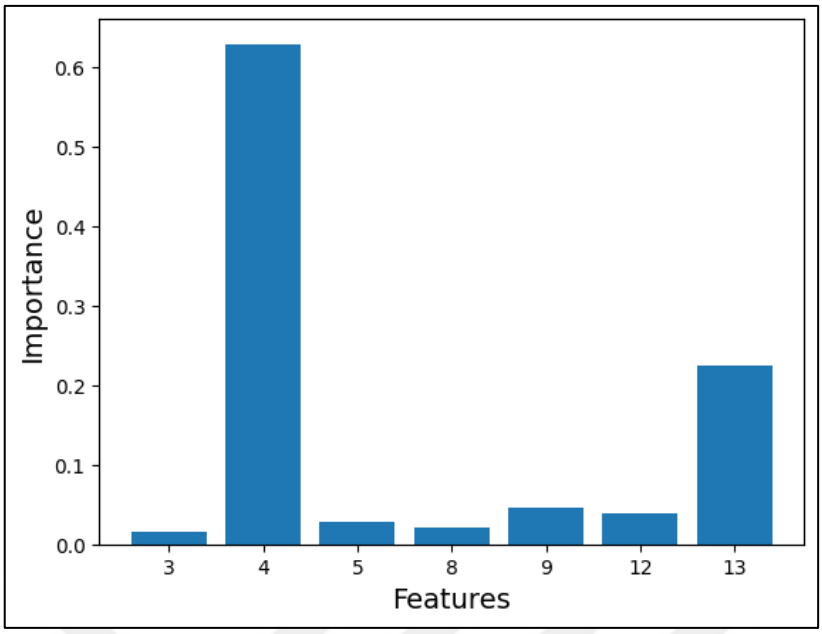

Figure 4.29: Feature Importance Analysis for Switzerland Dataset with RFE.

By employing RFE, the ANFIS model could focus on the most telling features, enhancing its predictive performance while offering insights into the critical factors that warrant attention for cardiovascular disease risk assessment. This form of feature importance analysis is not just a technical step in model optimization; it also provides valuable knowledge that can inform clinical decision-making and personalized patient care.

4.3.3 Performance Metrics and Results

The ANFIS model underwent a rigorous performance evaluation, demonstrating its capability to predict cardiovascular diseases across several datasets. The results, outlined below, highlight the model's diagnostic accuracy and reliability.

For the Cleveland dataset, the ANFIS model delivered an exemplary performance, flawlessly categorizing all instances with accuracy, precision, and recall of 100.00%. This impeccable classification is captured in the F1 score, which achieved a perfect 1.00, and the F2 score, equally reaching 1.00, reflecting the model's exceptional balance in predictive accuracy.

The model's superior performance is further substantiated by the ROC-AUC score, which stands at a perfect 1.00. This score, along with the ROC curve showcased in the provided Figure 4.30, signifies the model's unparalleled ability to differentiate between classes with no observable error.

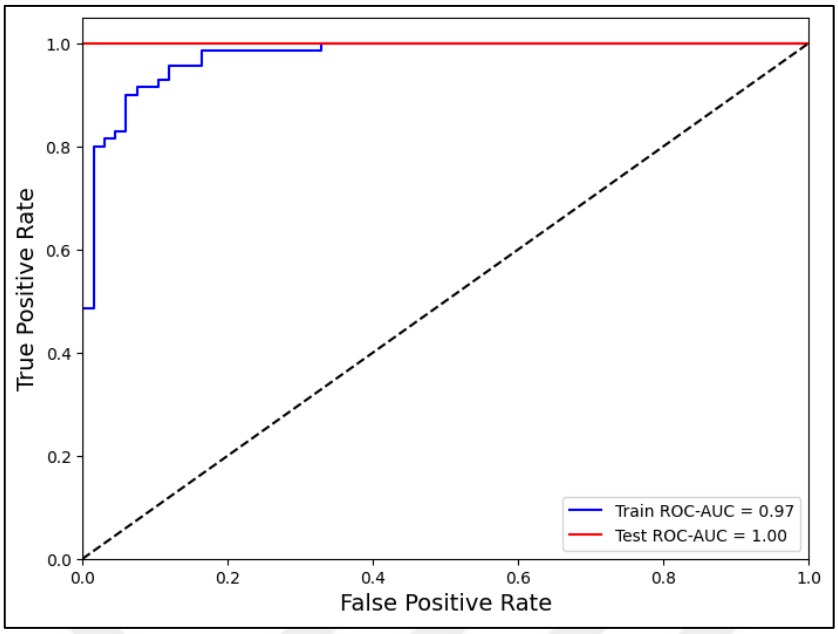

Figure 4.30: ROC Curve for Cleveland Dataset in ANFIS.

The accompanying confusion matrix Figure 4.31, displaying complete agreement bet confirm true labels and the model's predictions, with 29 true negatives and 30 true positives, confirms the absence of false classifications. This level of performance illustrates the model's precision in clinical diagnostics, demonstrating its potential as a reliable tool for cardiovascular disease detection.

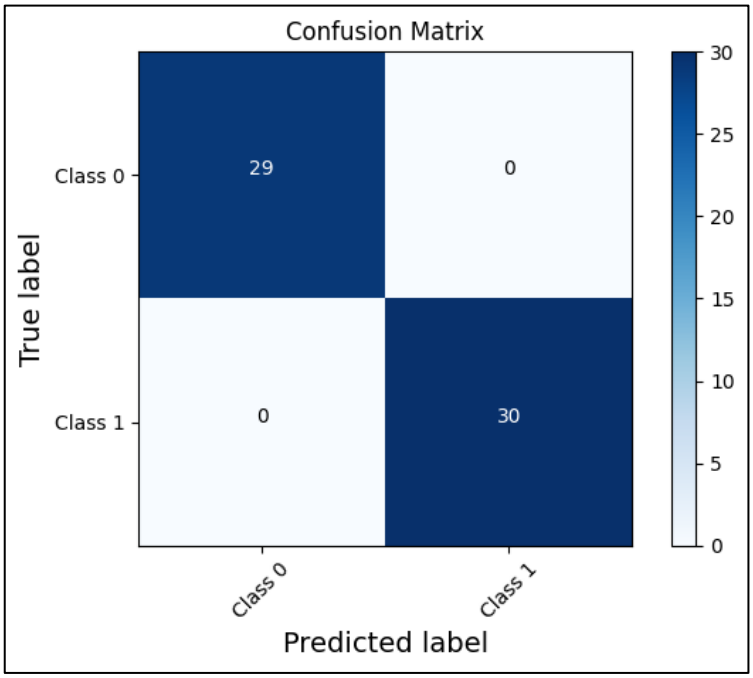

Figure 4.31: Confusion Matrix for Cleveland Dataset in ANFIS.

In analyzing the Hungarian dataset, the ANFIS model showcased commendable diagnostic accuracy with a performance of 96%. The precision and recall metrics both reflect a value of 0.96, culminating in an F1 score that maintains the same level of precision and sensitivity. Slightly more weight was given to recall, as evident from an F2 score of 0.97, reinforcing the model's focus on minimizing false negatives—a critical consideration in medical diagnosis.

The (ROC) curve and the corresponding ROC-AUC score of 0.96 provide a visual and statistical confirmation of the model's effective classification abilities. The close proximity of the test ROC curve to the upper left corner, as depicted in the ROC chart Figure 4.32, is indicative of the high true positive rate and low false positive rate, hallmarking a reliable predictive model.

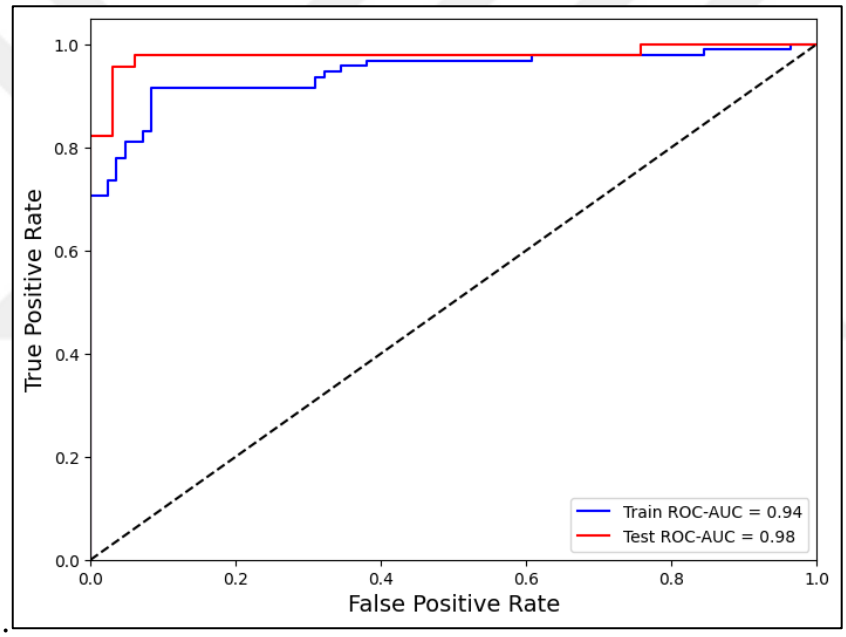

Figure 4.32: ROC Curve for Hungarian Dataset in ANFIS.

Supporting this, the confusion matrix Figure 4.33 displays a near-perfect classification with 31 true negatives and 44 true positives, suggesting that the model can distinguish between the absence and presence of the disease with high accuracy.

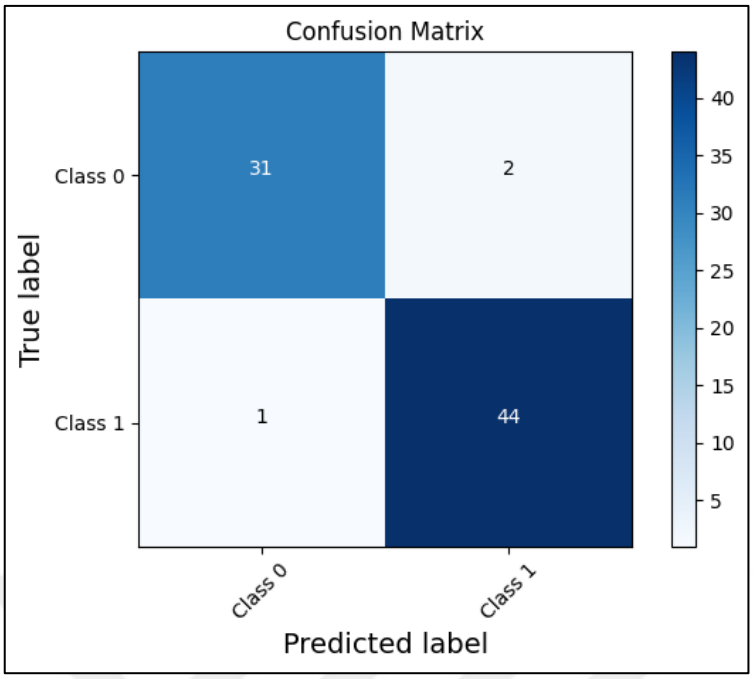

Figure 4.33: Confusion Matrix for Hungarian Dataset in ANFIS.

When scrutinizing the Long Beach VA dataset, the ANFIS model displayed a commendable level of accuracy at 78%. The precision, at 0.84, indicates a high rate of true positive predictions among all positive calls made by the model, while the recall rate of 0.78 assures that a substantial proportion of actual positive cases were correctly identified. These figures converge to produce an F1 score of 0.78, signifying a balanced harmonic mean between precision and recall.

A more nuanced look at the F2 score, which leans towards recall, shows a slightly higher value of 0.87, highlighting the model's capacity to correctly identify true positive cases, which is often more crucial in medical diagnostics. The ROC-AUC score further demonstrates the model's performance, with a solid ROC-AUC of 0.81, as illustrated in the ROC chart Figure 4.34.

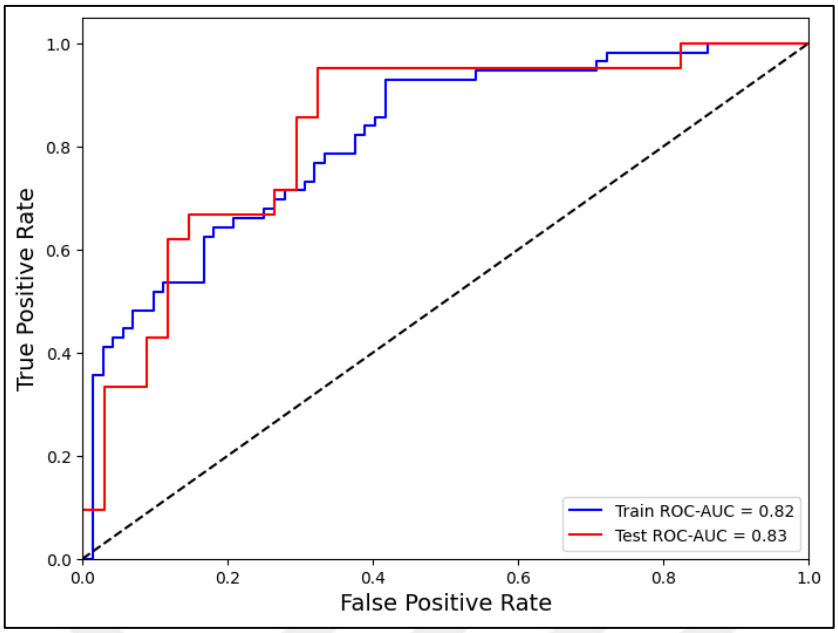

Figure 4.34: ROC Curve for Long Beach VA Dataset in ANFIS.

The confusion matrix Figure 4.35 for the Long Beach VA dataset reinforces these statistics, with 23 true negatives and 20 true positives. The presence of 11 false positives and only 1 false negative sheds light on the model's tendency to avoid false negatives at the expense of a slightly increased false positive rate.

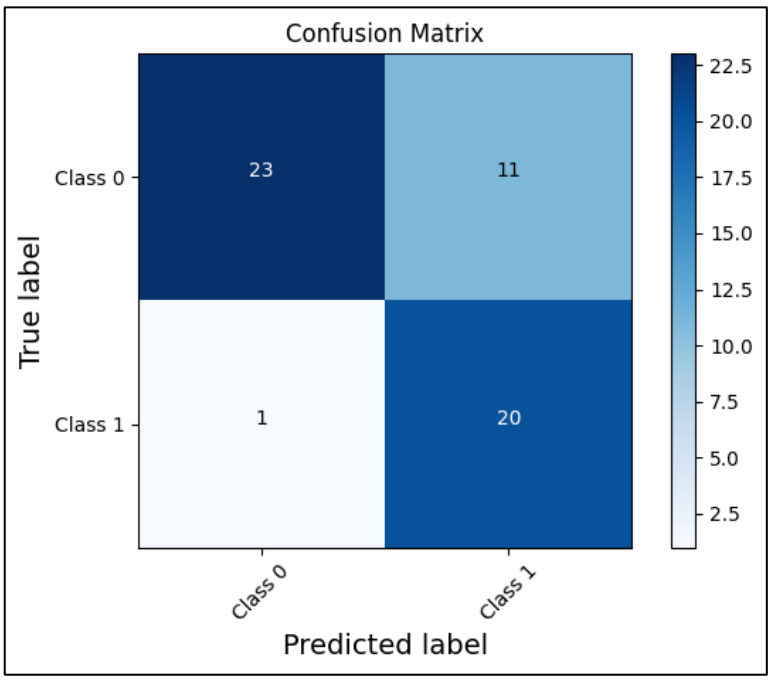

Figure 4.35: Confusion Matrix for Long Beach VA Dataset in ANFIS.

Upon evaluating the Switzerland dataset, the ANFIS model turned in a laudable performance with a 97% accuracy rate. This high level of accuracy is matched by a precision and recall of 0.97, denoting that the model correctly identified most of both positive and negative cases. The F1 score, which is a measure of the model's accuracy considering both precision and recall, stands strong at 0.97. The F2 score, which places more emphasis on recall, is slightly higher at 0.98, suggesting that the model is particularly effective at capturing true positive cases.

The ROC-AUC score is an exceptional 0.97, which indicates a high ability of the model to differentiate between the classes effectively. This is visually confirmed by the ROC curve presented in Figure 4.36, with the test ROC-AUC curve nearing the top-left corner, signifying an excellent true positive rate with minimal false positives.

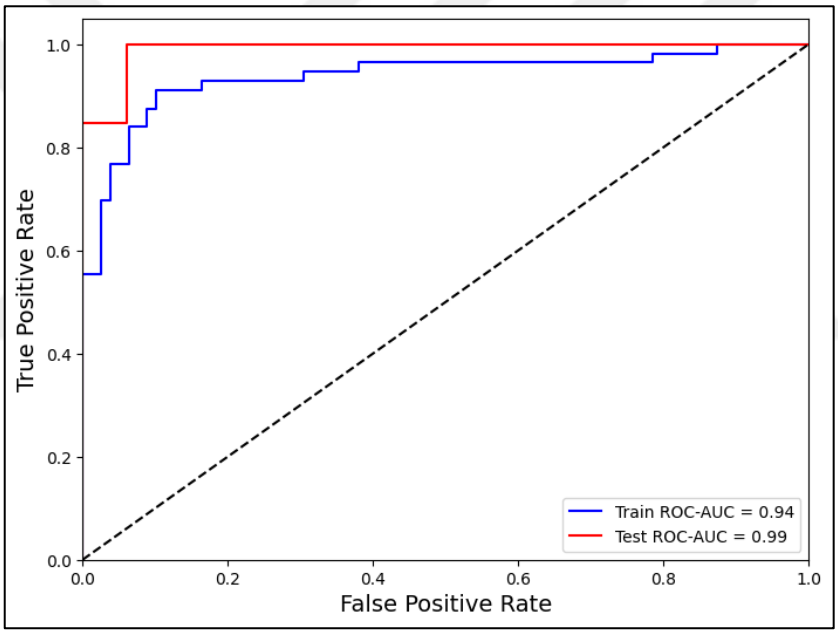

Figure 4.36: ROC Curve for Switzerland Dataset in ANFIS.

The confusion matrix Figure 4.37 further illustrates the model's high level of correct classifications, with 31 true negatives and 26 true positives, against a mere 2 false positives, highlighting the model's precision in detecting cardiovascular disease.

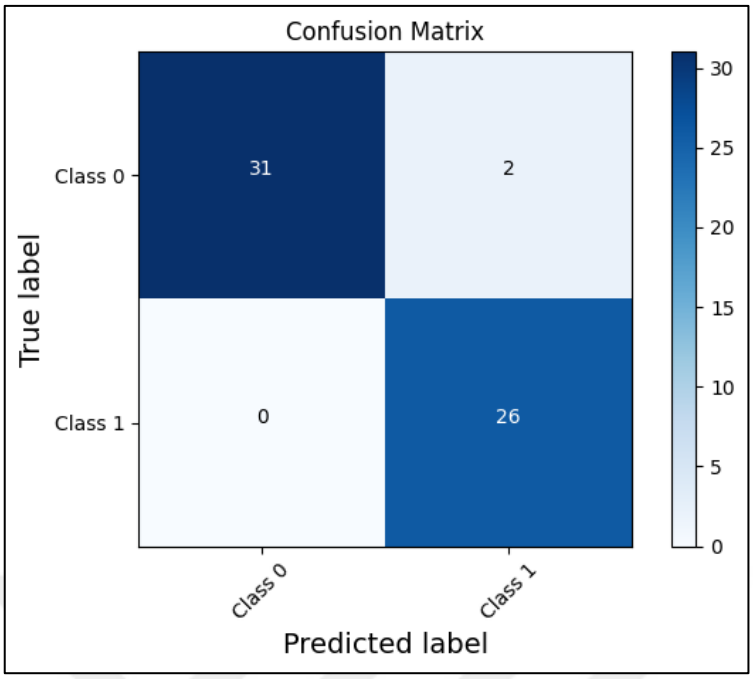

Figure 4.37: Confusion Matrix for Switzerland Dataset in ANFIS.

For the Statlog Heart dataset, the ANFIS model's performance was solid, achieving 95% accuracy. The model showed a balanced precision and recall, each at 0.95, which led to an F1 score of the same value. The F2 score was elevated to 0.98, reflecting the model’s strong emphasis on recall, thus prioritizing the correct identification of all positive cases.

The ROC-AUC score, which quantifies the model's ability to differentiate between classes, matched the accuracy at a robust 0.95. This suggests a highly effective model with a commendable true positive rate and a low false positive rate, as evidenced by the ROC curve in Figure 4.38.

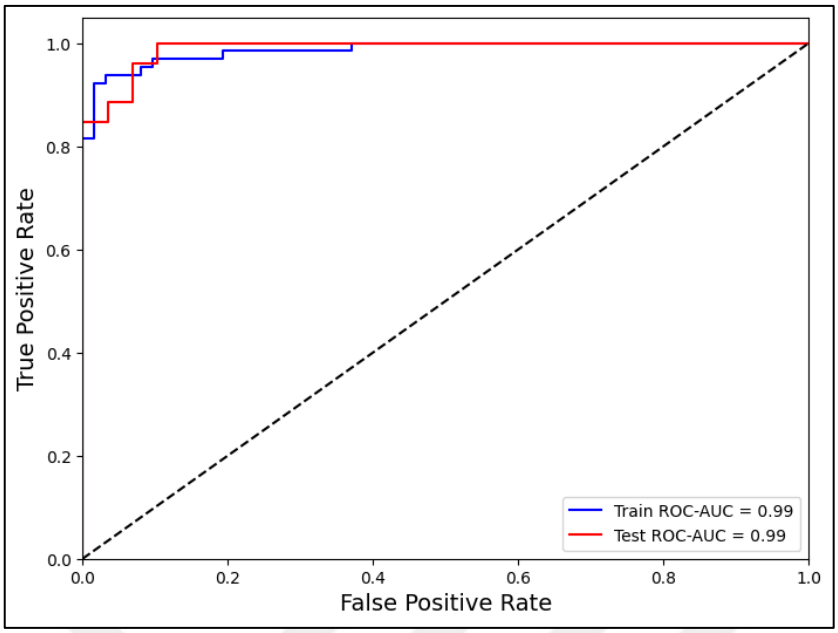

Figure 4.38: ROC Curve for Statlog Heart Dataset in ANFIS.

The confusion matrix Figure 4.39 reinforces this performance, indicating 26 true positives and 26 true negatives, with just a minor number of 3 false positives, demonstrating the model's capability to accurately classify the instances.

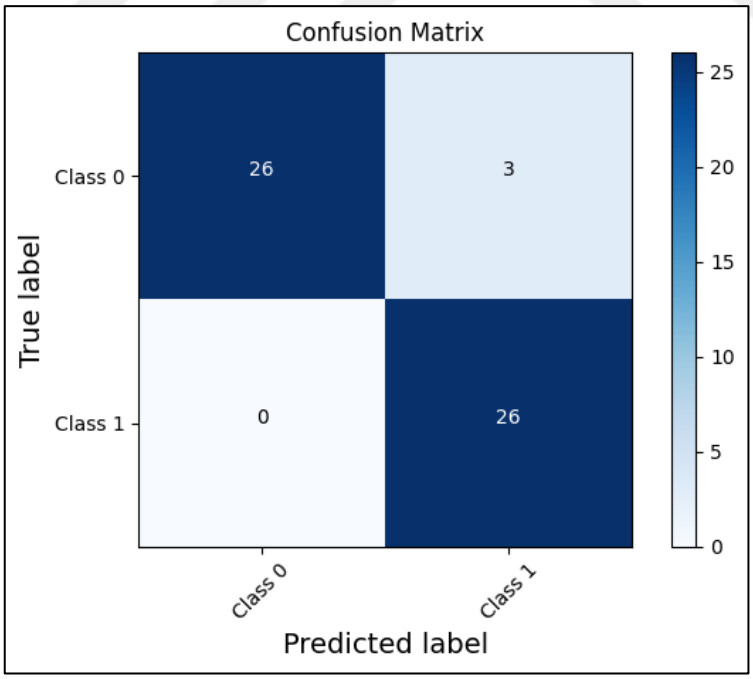

Figure 4.39: Confusion Matrix for Statlog Heart Dataset in ANFIS.

The ANFIS model's performance across the various datasets underscores its reliability and effectiveness in the diagnosis of cardiovascular diseases. Despite the natural variability in results that comes with different data sources, the model has demonstrated a consistently high degree of precision and recall, which are critical for clinical diagnostic tools. The favorable F2 scores indicate a thoughtful consideration of recall, suggesting that the model is well-tuned to recognize true positive cases effectively which is a crucial aspect of medical diagnostics. Moreover, the ROC-AUC scores across all datasets affirm the model's discriminating power, adding to its credibility as an asset in healthcare settings for the prediction and analysis of cardiovascular conditions.

4.4 COMPARATIVE ANALYSIS OF XGBOOST AND ANFIS MODEL

This section delves into the comparative performance of the XGBoost and ANFIS models, employing a holistic view of their outcomes across multiple datasets as outlined in Table 4.2 and Table 4.3 below.

Table 4.2: XGBoost Model Results.

Datasets	Accuracy	Precision	Recall	F1 Score	F2 Score	ROC-AUC Score
Cleveland	98.28%	0.97	1.00	0.98	0.99	0.98
Hungarian	99%	0.98	1.00	0.99	1.00	0.98
Long-beach-va	97%	0.96	0.96	0.96	0.96	0.96
Switzerland	100.00%	1.00	1.00	1.00	1.00	1.00
Statlog_heart	100.00%	1.00	1.00	1.00	1.00	1.00

Table 4.3: ANFIS Model Results.

Datasets	Accuracy	Precision	Recall	F1 Score	F2 Score	ROC-AUC Score
Cleveland	100.00%	1.00	1.00	1.00	1.00	1.00
Hungarian	96%	0.96	0.96	0.96	0.97	0.96
Long-beach-va	78%	0.84	0.78	0.78	0.87	0.81
Switzerland	97%	0.97	0.97	0.97	0.98	0.97
Statlog_heart	95%	0.95	0.95	0.95	0.98	0.95

XGBoost, with its robustness and efficiency, excelled in most datasets, achieving near-perfect scores across all performance metrics. This suggests that XGBoost is highly capable of handling diverse, high-dimensional data, making it a formidable tool for cardiovascular disease prediction.

On the other hand, the ANFIS model, with its unique blend of fuzzy logic and neural network architecture, showed a remarkable ability to capture complex patterns. Although it presented with slight variability, it maintained commendable precision and recall levels, emphasizing its potential as a nuanced diagnostic tool.

When directly compared, the XGBoost model often edged out with slightly superior accuracy and consistency. However, the ANFIS model demonstrated a particular strength in datasets where the underlying data patterns were less linear and more intricate, which could be attributed to its fuzzy logic component.

The comparative analysis suggests that while XGBoost can be favored for its predictive power and robustness, ANFIS provides an alternative perspective with its ability to handle complexity and ambiguity in data, which are common in medical datasets. The choice between the two may ultimately depend on the specific dataset characteristics and the clinical context in which they are applied.

The insights from this comparative analysis highlight the importance of model selection in healthcare analytics. By understanding the unique advantages of each model, researchers and practitioners can make informed decisions to utilize the appropriate model that best aligns with the diagnostic task at hand, thereby improving patient outcomes and enhancing clinical decision-making.

4.5 CONTRIBUTION TO CARDIOVASCULAR DISEASE PREDICTION IN HEALTHCARE

This section delineates the contributions of our investigation within the expanding field of cardiovascular disease prediction, with a particular emphasis on the innovative advancements as presented in recent scholarly literature.

Our research augments the extant scholarly discourse by facilitating a detailed comparative analysis of the XGBoost and ANFIS models, outlined in our proposed model in Table 4.4. This comparison underscores their unique and combined strengths in forecasting cardiovascular ailments. Moreover, our evaluation extends beyond mere accuracy metrics to include precision, recall, and area under the Receiver Operating Characteristic (ROC) curve, thereby providing a holistic view of model efficacy.

When our findings are positioned alongside contemporary studies, the alignment with the forefront of research in this domain is apparent. For instance, in reference [52], the integration of Modified Salp Swarm Optimization with the Adaptive Neuro-Fuzzy Inference System (MSSO-ANFIS) within the Internet of Medical Things (IoMT) environment achieved an accuracy of 99.45%, as shown in Table 4.4. Our study mirrors this precision but also delves deeper into a nuanced analysis of performance metrics.

In parallel, reference [51] explores the deployment of an Enhanced Deep Learning Assisted Convolutional Neural Network (EDCNN) for cardiac disease detection on IoMT platforms, which achieved an accuracy of 99.1%, as indicated in Table 4.4. Our research contributes to the discourse by exploring alternative machine learning strategies that provide similar levels of performance without the exclusive reliance on deep learning frameworks.

Further, references [50] and [49] detail the application of systems employing feature fusion and ensemble deep learning, as well as diverse ML algorithms such as Random Forest and SVM, respectively, as detailed in Table 4.4. These methodologies resonate with our approach to assess and validate a spectrum of predictive techniques, thereby pushing the boundaries of intelligent healthcare systems.

The utilization of the High-Dimensional Partitioning Method (HDPM) highlighted in paper [48] and the combination of SVM with fuzzy logic in paper [47], both documented in Table 4.4, represent significant advancements in this sphere. Our study complements these methodologies by contrasting the robust, tree-based XGBoost model with the adaptable,

rule-based ANFIS model, thus enriching the toolkit available for clinical decision support systems.

Through a systematic evaluation of the strengths and applications of both XGBoost and ANFIS models, our study not only corroborates the high accuracy levels cited in the literature but also enriches the decision-making processes with in-depth performance metrics. These insights are pivotal for the development of more refined and adaptable predictive systems, which address the intricacies of cardiovascular disease diagnosis and pave the way for more individualized and precise patient care within the IoMT framework.

Table 4.4: Comparative Overview of Proposed and Existing Studies.

References	Objective	ML Technique	Datasets	Accuracy
[52]	Enhance heart disease prediction accuracy in IoMT	MSSO-ANFIS	Hungarian	99.45%
			Framingham	
[51]	Develop EDCNN for heart disease detection on IoMT	EDCNN with feature extraction and Bayesian classification.	UCI repository dataset	99.1%
[50]	Create a smart system for heart disease prediction	Feature fusion, information gain, conditional probability, ensemble deep learning	Cleveland	98.5%
			Hungarian	
[49]	Develop IoT-based healthcare monitoring	Random Forest, K-NN, SVM, Decision Trees, MLP.	Several public datasets from the UCI ML Repository	97.26%
[48]	Develop HDPM for clinical decision support system	DBSCAN, SMOTE-ENN, XGBoost.	Cleveland	98.4%
			Statlog	95.90%
[47]	Develop SVM with fuzzy logic for heart prediction	SVM, fuzzy decision fusion	Cleveland	96.23%
Proposed model	Develop a prediction model using machine learning algorithms	XGBoost	Cleveland	98.28%
			Hungarian	99%
			Long-beach-va	97%
			Switzerland	100.00%
			Statlog_heart	100.00%
		ANFIS	Cleveland	100.00%
			Hungarian	96%
			Long-beach-va	78%
			Switzerland	97%
			Statlog_heart	95%

This tabular synthesis articulates the scope and findings from various investigations, juxtaposing them with our research to underscore the significant contributions our work makes to the burgeoning field of cardiovascular disease prediction using advanced computational techniques within the IoMT landscape.

5. CONCLUSION

5.1 OVERVIEW OF RESEARCH

This thesis conducted on a comprehensive journey to explore the capabilities of XGBoost and ANFIS machine learning models in predicting cardiovascular diseases. The foray into these sophisticated algorithms was driven by the pressing need to enhance diagnostic accuracy and reliability in the face of the global burden of cardiovascular diseases.

5.2 SUMMARY OF FINDINGS

Our in-depth analysis revealed that both XGBoost and ANFIS models exhibit remarkable proficiency in classifying heart disease. By meticulously processing five distinct datasets, the study not only reinforced the models' versatility but also their adeptness at handling diverse patient data. The XGBoost model demonstrated exceptional performance, particularly in the Switzerland and Statlog datasets, achieving a perfect accuracy score of 100%. Similarly, the ANFIS model shone with a flawless accuracy rate in the Cleveland dataset and strong results across others.

5.3 COMPARATIVE ANALYSIS

A comparative evaluation between the two models highlighted their individual strengths. The XGBoost model's robustness is evident in its consistent ability to balance precision and recall, substantiated by high F1 and F2 scores. On the other hand, the ANFIS model, with its integration of fuzzy logic, showcased a compelling capacity to model the nonlinear and complex patterns often present in medical diagnosis.

5.4 IMPLICATIONS AND CONTRIBUTIONS

Our work contributes significant new insights to the field of medical diagnostics. It serves as a testament to the potential of integrating advanced machine learning techniques into the healthcare domain. The findings of this thesis reinforce the narrative that machine learning models like XGBoost and ANFIS can be transformative tools in disease prediction, especially within the IoMT framework.

5.5 RECOMMENDATIONS AND FUTURE WORK

While this study has laid a robust foundation, future research could expand upon this work by integrating larger and more varied datasets, potentially including real-time patient data, to further validate and refine the models. Additionally, exploring the synergy of hybrid models combining the strengths of XGBoost and ANFIS could yield even more powerful predictive tools.

In conclusion, the results of this thesis underscore the viability of machine learning applications in healthcare, particularly for heart disease prediction. It demonstrates a clear path forward for the integration of algorithms like XGBoost and ANFIS in medical diagnostics, which can offer enhanced precision and potentially personalized patient care strategies. While there are still challenges to be met, such as data diversity and model interpretability, the progress made suggests that with further validation and refinement, machine learning could significantly improve the efficiency and effectiveness of disease diagnosis. This research contributes to the ongoing transformation of health data analytics, bringing us closer to a future where healthcare is increasingly informed by intelligent data-driven decisions.

REFERENCES

[1] "Cardiovascular diseases (CVDs)." Accessed: Mar. 08, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] B. Bozkurt *et al.*, "Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure: Endorsed by the Canadian Heart Failure Society, Heart Failure Association of India, Cardiac Society of Australia and New Zealand, and Chinese Heart Failure Association," *Eur J Heart Fail*, vol. 23, no. 3, pp. 352–380, Mar. 2021, doi: 10.1002/ejhf.2115.

[3] C. W. Tsao *et al.*, "Heart Disease and Stroke Statistics - 2023 Update: A Report from the American Heart Association," *Circulation*, vol. 147, no. 8. Lippincott Williams and Wilkins, pp. E93–E621, Feb. 21, 2023. doi: 10.1161/CIR.0000000000001123.

[4] D. K. Arnett *et al.*, "2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines," *Circulation*, vol. 140, no. 11. NLM (Medline), pp. e596–e646, Sep. 10, 2019. doi: 10.1161/CIR.0000000000000678.

[5] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," *Future Generation Computer Systems*, vol. 111, pp. 714–722, Oct. 2020, doi: 10.1016/j.future.2019.09.056.

[6] P. Ponikowski *et al.*, "Heart failure: preventing disease and death worldwide," *ESC Heart Failure*, vol. 1, no. 1. Wiley-Blackwell, pp. 4–25, Sep. 01, 2014. doi: 10.1002/ehf2.12005.

[7] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[8] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, “Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,” *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[9] E. J. Benjamin *et al.*, “Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association,” *Circulation*, vol. 139, no. 10, pp. e56–e528, Mar. 2019, doi: 10.1161/CIR.0000000000000659.

[10] W. Dai, T. S. Brisimi, W. G. Adams, T. Mela, V. Saligrama, and I. C. Paschalidis, “Prediction of hospitalization due to heart diseases by supervised learning methods,” *Int J Med Inform*, vol. 84, no. 3, pp. 189–197, Mar. 2015, doi: 10.1016/j.ijmedinf.2014.10.002.

[11] A. Garg and V. Mago, “Role of machine learning in medical research: A survey,” *Computer Science Review*, vol. 40. Elsevier Ireland Ltd, May 01, 2021. doi: 10.1016/j.cosrev.2021.100370.

[12] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, “An Automated Diagnostic System for Heart Disease Prediction Based on $\chi 2$ Statistical Model and Optimally Configured Deep Neural Network,” *IEEE Access*, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

[13] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, “An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,” *Expert Syst Appl*, vol. 68, pp. 163–172, Feb. 2017, doi: 10.1016/j.eswa.2016.10.020.

[14] S. S. Sarmah, “An Efficient IoT-Based Patient Monitoring and Heart Disease Prediction System Using Deep Learning Modified Neural Network,” *IEEE Access*, vol. 8, pp. 135784–135797, 2020, doi: 10.1109/ACCESS.2020.3007561.

[15] M. Vaduganathan, G. A. Mensah, J. V. Turco, V. Fuster, and G. A. Roth, “The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health,” *Journal of the American College of Cardiology*, vol. 80, no. 25. Elsevier Inc., pp. 2361–2371, Dec. 20, 2022. doi: 10.1016/j.jacc.2022.11.005.

[16] B. Ziaeian and G. C. Fonarow, “Epidemiology and aetiology of heart failure,” *Nature Reviews Cardiology*, vol. 13, no. 6. Nature Publishing Group, pp. 368–378, Jun. 01, 2016. doi: 10.1038/nrcardio.2016.25.

[17] “How Blood Flows Through the Heart & Body.” Accessed: Mar. 12, 2024. [Online]. Available: https://my.clevelandclinic.org/health/articles/17060-how-does-the-blood-flow-through-your-heart

[18] “Report on Cardiovascular Health and Diseases in China 2021: An Updated Summary,” *Biomedical and Environmental Sciences*, vol. 35, no. 7, pp. 573–603, Jul. 2022, doi: 10.3967/bes2022.079.

[19] Institute of Electrical and Electronics Engineers, *2017 IEEE Symposium on Computers and Communications (ISCC).*

[20] L. Ali *et al.*, “An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure,” *IEEE Access*, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

[21] C. Albus, J. Barkhausen, E. Fleck, J. Haasenritter, O. Lindner, and S. Silber, “The Diagnosis of Chronic Coronary Heart Disease,” *Deutsches Arzteblatt international*, vol. 114, no. 42. pp. 712–719, Oct. 20, 2017. doi: 10.3238/arztebl.2017.0712.

[22] “Cardiovascular Disease: Types, Causes & Symptoms.” Accessed: Mar. 12, 2024. [Online]. Available: https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease

[23] “Diagnosing a Heart Attack | American Heart Association.” Accessed: Mar. 12, 2024. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/diagnosing-a-heart-attack

[24] “Cardiovascular Imaging Section | Cleveland Clinic.” Accessed: Mar. 12, 2024. [Online]. Available: https://my.clevelandclinic.org/departments/heart/depts/cardiovascular-imaging.

[25] R. R. Sun, M. Liu, L. Lu, Y. Zheng, and P. Zhang, “Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments,” *Cell Biochem Biophys*, vol. 72, no. 3, pp. 857–860, Jul. 2015, doi: 10.1007/s12013-015-0551-6.

[26] “Exercise Stress Test | American Heart Association.” Accessed: Mar. 12, 2024. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/diagnosing-a-heart-attack/exercise-stress-test

[27] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, “An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection,” *IEEE Access*, vol. 7, pp. 180235–180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

[28] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, “Predicting the Risk of Heart Failure with EHR Sequential Data Modeling,” *IEEE Access*, vol. 6, pp. 9256–9261, Jan. 2018, doi: 10.1109/ACCESS.2017.2789324.

[29] H. Habehh and S. Gohel, “Machine Learning in Healthcare,” *Curr Genomics*, vol. 22, no. 4, pp. 291–300, Jul. 2021, doi: 10.2174/1389202922666210705124359.

[30] Q. An, S. Rahman, J. Zhou, and J. J. Kang, “A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges,” *Sensors*, vol. 23, no. 9. MDPI, May 01, 2023. doi: 10.3390/s23094178.

[31] R. Bhardwaj, A. R. Nambiar, and D. Dutta, “A Study of Machine Learning in Healthcare,” in *Proceedings - International Computer Software and Applications Conference*, IEEE Computer Society, Sep. 2017, pp. 236–241. doi: 10.1109/COMPSAC.2017.164.

[32] IEEE Staff, *2019 Amity International Conference on Artificial Intelligence (AICAI)*. IEEE, 2019.

[33] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” *IEEE Access*, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[34] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, “Systematic poisoning attacks on and defenses for machine learning in healthcare,” *IEEE J Biomed Health Inform*, vol. 19, no. 6, pp. 1893–1905, Nov. 2015, doi: 10.1109/JBHI.2014.2344095.

[35] P. Karmani, A. A. Chandio, I. A. Korejo, and M. S. Chandio, “A Review of Machine Learning for Healthcare Informatics Specifically Tuberculosis Disease Diagnostics,” in *Communications in Computer and Information Science*, Springer Verlag, 2019, pp. 50–61. doi: 10.1007/978-981-13-6052-7_5.

[36] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, “Ethical Machine Learning in Healthcare,” *Annual Review of Biomedical Data Science Annu. Rev. Biomed. Data Sci*, vol. 2021, pp. 123–144, 2021, doi: 10.1146/annurev-biodatasci-092820.

[37] A. Alanazi, “Using machine learning for healthcare challenges and opportunities,” *Informatics in Medicine Unlocked*, vol. 30. Elsevier Ltd, Jan. 01, 2022. doi: 10.1016/j.imu.2022.100924.

[38] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars and applications,” *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, Jan. 2022, doi: 10.1016/j.ijin.2022.05.002.

[39] T. Vivekanandan and N. C. Sriman Narayana Iyengar, “Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease,” *Comput Biol Med*, vol. 90, pp. 125–136, Nov. 2017, doi: 10.1016/j.compbiomed.2017.09.011.

[40] A. K. Dwivedi, “Performance evaluation of different machine learning techniques for prediction of heart disease,” *Neural Comput Appl*, vol. 29, no. 10, pp. 685–693, May 2018, doi: 10.1007/s00521-016-2604-1.

[41] S. M. Saqlain *et al.*, “Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines,” *Knowl Inf Syst*, vol. 58, no. 1, pp. 139–167, Jan. 2019, doi: 10.1007/s10115-018-1185-y.

[42] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, vol. 36, pp. 82–93, Mar. 2019, doi: 10.1016/j.tele.2018.11.007.

[43] P. Ghosh *et al.*, “Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques,” *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[44] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[45] K. Budholiya, S. K. Shrivastava, and V. Sharma, “An optimized XGBoost based diagnostic system for effective prediction of heart disease,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.

[46] A. Gupta, R. Kumar, H. Singh Arora, and B. Raman, “MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis,” *IEEE Access*, vol. 8, pp. 14659–14674, 2020, doi: 10.1109/ACCESS.2019.2962755.

[47] M. W. Nadeem, H. G. Goh, M. A. Khan, M. Hussain, M. F. Mushtaq, and V. A. P. Ponnusamy, “Fusion-Based Machine Learning Architecture for Heart Disease Prediction,” *Computers, Materials and Continua*, vol. 67, no. 2, pp. 2481–2496, 2021, doi: 10.32604/cmc.2021.014649.

[48] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, “HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System,” *IEEE Access*, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[49] P. Kaur, R. Kumar, and M. Kumar, “A healthcare monitoring system using random forest and internet of things (IoT),” *Multimed Tools Appl*, vol. 78, no. 14, pp. 19905–19916, Jul. 2019, doi: 10.1007/s11042-019-7327-8.

[50] F. Ali *et al.*, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Information Fusion*, vol. 63, pp. 208–222, Nov. 2020, doi: 10.1016/j.inffus.2020.06.008.

[51] Y. Pan, M. Fu, B. Cheng, X. Tao, and J. Guo, "Enhanced deep learning assisted convolutional neural network for heart disease prediction on the internet of medical things platform," *IEEE Access*, vol. 8, pp. 189503–189512, 2020, doi: 10.1109/ACCESS.2020.3026214.

[52] M. A. Khan and F. Algarni, "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," *IEEE Access*, vol. 8, pp. 122259–122269, 2020, doi: 10.1109/ACCESS.2020.3006424.

[53] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[54] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information (Switzerland)*, vol. 13, no. 10, Oct. 2022, doi: 10.3390/info13100475.

[55] M. A. Denai, F. Palis, and A. Zeghbib, "ANFIS Based Modelling and Control of Non-linear Systems : A tutorial *."

[56] J. Sekar, P. Aruchamy, H. Sulaima Lebbe Abdul, A. S. Mohammed, and S. Khamuruddeen, "An efficient clinical support system for heart disease prediction using TANFIS classifier," *Comput Intell*, vol. 38, no. 2, pp. 610–640, Apr. 2022, doi: 10.1111/coin.12487.

[57] D. Karaboga and E. Kaya, "Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey," *Artificial Intelligence Review*, vol. 52, no. 4. Springer Netherlands, pp. 2263–2293, Dec. 01, 2019. doi: 10.1007/s10462-017-9610-2.

[58] S. W. P. M. Janosi Andras and R. Detrano, "Heart Disease." 1988.

[59] “Statlog (Heart) - UCI Machine Learning Repository.” Accessed: Apr. 02, 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/145/statlog+heart

[60] S. García, J. Luengo, and F. Herrera, “Tutorial on practical tips of the most influential data preprocessing algorithms in data mining,” *Knowl Based Syst*, vol. 98, pp. 1–29, Apr. 2016, doi: 10.1016/j.knosys.2015.12.006.

[61] IEEE Staff, *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE, 2019.

[62] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” *Mobile Information Systems*, vol. 2018, 2018, doi: 10.1155/2018/3860146.

[63] C. A. Owen, G. Dick, and P. A. Whigham, “Standardization and Data Augmentation in Genetic Programming,” *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 6, pp. 1596–1608, Dec. 2022, doi: 10.1109/TEVC.2022.3160414.

[64] T. Al-shehari and R. A. Alsowail, “An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques,” *Entropy*, vol. 23, no. 10, Oct. 2021, doi: 10.3390/e23101258.

[65] Global IT Research Institute, IEEE Communications Society, and Institute of Electrical and Electronics Engineers, *The IEEE 20th International Conference on Advanced Communications Technology : “Opening New Era of Intelligent Things!” : ICACT 2018 : Elysian Gangchon, Chuncheon, Korea (South) : Feb. 11-14, 2018 : proceeding & journal*.

[66] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data.”

[67] “SMOTEENN — Version 0.12.0.” Accessed: Mar. 17, 2024. [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.combine.SMOTEENN.html#imblearn.combine.SMOTEENN

[68] E. Martel *et al.*, “Implementation of the Principal Component Analysis onto high-performance computer facilities for hyperspectral dimensionality reduction: Results and comparisons,” *Remote Sens (Basel)*, vol. 10, no. 6, Jun. 2018, doi: 10.3390/rs10060864.

[69] “sklearn.decomposition.PCA — scikit-learn 1.4.1 documentation.” Accessed: Mar. 17, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#

[70] M. Kuhn and K. Johnson, “Feature Engineering and Selection; A Practical Approach for Predictive Models; Edition 1.” doi: 10.4324/9781315108230.

[71] “sklearn.feature_selection.SelectFromModel — scikit-learn 1.4.1 documentation.” Accessed: Mar. 17, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html#

[72] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[73] F. Jabeen *et al.*, “An IoT based efficient hybrid recommender system for cardiovascular disease,” *Peer Peer Netw Appl*, vol. 12, no. 5, pp. 1263–1276, Sep. 2019, doi: 10.1007/s12083-019-00733-3.

[74] A. Baccouche, B. Garcia-Zapirain, C. C. Olea, and A. Elmaghraby, “Ensemble deep learning models for heart disease classification: A case study from Mexico,” *Information (Switzerland)*, vol. 11, no. 4, Apr. 2020, doi: 10.3390/INFO11040207.

[75] G. S. Fuhnwi, M. Revelle, and C. Izurieta, “Improving Network Intrusion Detection Performance : An Empirical Evaluation Using Extreme Gradient Boosting (XGBoost) with Recursive Feature Elimination,” in *2024 IEEE 3rd International Conference on AI in Cybersecurity, ICAIC 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICAIC60265.2024.10433805.

[76] “Machine Learning Theory and Applications: Hands-on Use Cases with Python on ... - Xavier Vasques - Google Books.” Accessed: Apr. 03, 2024. [Online]. Available: https://books.google.com.tr/books?id=BrvtEAAAQBAJ&pg=PR3&lpg=PR3&dq=Machine+Learning+Theory+and+Applications%0D%0AHands-on+Use+Cases+with+Python+on+Classical+and+Quantum+Machines%0D%0AXavier+Vasques%0D%0AIBM+Technology,+Bois-Colombes,+France%0D%0ALaboratoire+de+Recherche+en+Neurosciences+Cliniques,+Montferriez+sur+lez,+France%0D%0AEcole+Nationale+Sup%C3%A9rieure+de+Cognitique+Bordeaux,+Bordeaux,+France+cite&source=bl&ots=EJR7ntIa7_&sig=ACfU3U2fAABjQoRbLKXfqhMA0mX37QaKAQ&hl=en&sa=X&ved=2ahUKEwiG0onU2aSFAxWKQ_EDHWC4AQAQ6AF6BAgUEAM#v=onepage&q&f=false